

Translating and the Computer 45



20-22 November 2023

European Convention Center, Luxembourg

Proceedings



ISBN 978-2-9701733-1-1



November 2024. Editions Tradulex, Geneva

© AsLing, The International Association for Advancement in Language Technology

This document is downloadable from www.tradulex.com and www.asling.org

Acknowledgements

AsLing wishes to thank and acknowledge the support of the sponsors of TC45:

Gold Sponsors



Silver Sponsors



Bronze Sponsors



The Executive Committee of AsLing establishes several bodies each year, to organise and carry out the annual conference. Membership in these bodies overlap. The tables below show membership in these bodies for TC45.

Conference Organising Committee:

Vicent Briva-Iglesias, Dublin City University
Denis Dechandon, Publications Office of the European Union
Emanuelle Esperança-Rodier, Université de Grenoble-Alpes
Valentini Kalfadopoulos, Ionian University
Ruslan Mitkov, Lancaster University
Maria Recort Ruiz, International Labour Office
Lionel Shen, International Maritime Organization
Vilemini Sosoni, Ionian University
Olaf-Michael Stefanov, United Nations (ret.)
Nelson Verástegui, International Telecommunication Union (ret.)
Coordinator: Denis Dechandon

Editors of the Proceedings:

Amal Haddad Haddad
Vilemini Sosoni

Programme Committee:

Juan José Arevalillo, Hermes Traducciones
Lynne Bowker, University of Ottawa
Vicent Briva-Iglesias, Dublin City University
Sheila Castilho, Dublin City University/ADAPT Centre
Flix do Carmo, University of Surrey
David Chambers, AsLing Honorary Member
Félix do Carmo, University of Surrey
Gloria Corpas Pastor, University of Malaga
Ayten Dersan, World Trade Organization
Jorge Diaz-Cintas, Centre for Translation Studies (CenTraS) & University College London
Gökhan Doğru, Universitat Autònoma de Barcelona & Dublin City University
Joanna Drugan, Heriot-Watt University
Emmanuelle Esperança-Rodier, University of Grenoble Alps
María Fernandez-Parra, Swansea University
David Filip, Huawei Ireland Research Center
Amal Haddad Haddad, University of Granada
Valentini Kalfadopoulos, Ionian University
Elizabeth Marshman, University of Ottawa
Johanna Monti, L'Orientale University of Naples
Joss Moorkens, Dublin City University/ADAPT Centre
Antoni Oliver González, Universitat Oberta de Catalunya
Constantin Orăsan, University of Surrey
Michail Panagopoulos, Ionian University
Rozane Rebechi, University of Rio Grande do Sul
Celia Rico Perez, Universidad Complutense de Madrid
Vilemini Sosoni, Ionian University
Paola Valli, Tamedia
Nelson Verástegui, International Telecommunications Union (ret.)
Michał Ziemiński, World Intellectual Property Organization

Contents

New approaches to studying the cognitive impact of a CAI tool on Chinese interpreting trainees Du Zhiqiang and Ricardo Muñoz Martín	7
Poster on ChatGPT Translator Plus Terence Lewis	21
Large Language Models – origin, growth and capabilities Andrzej Zydroń, Rafał Jaworski and Szymon Kaczmarek	24
Envisioning the Post-Editor’s Workstation: A Backward Glance and a Glimpse into the Future Marie Escribe and Miguel Ángel Candel-Mora	37
Post-Editing Machine Translation Beyond the Binary: Insights into Gender Bias and Screen Activity Manuel Lardelli	50
The use of speech technologies and machine translation in institutional translation practices Justus Brockmann, Alina Secară and Dragoş Ciobanu	65
How can Paidiom improve the neural machine translation of multiword expressions? Carlos Manuel Hidalgo-Tenero and Francisco Javier Lima-Florida	81
Correcting biased translations with the Fairslator API Michal Měchura	98
Human & Machine Translation Quality: Comparing & Contrasting Concepts Bettina Hiebl and Dagmar Gromann	108
Subtitling videos within a language service: a hands-on approach Samuel Urscheler and Sandra Casas	129
From shifting thoughts to unlocking knowledge: The power of terminology in the digital era Denis Dechandon, Carolina Dunaevsky, Anikó Gerencsér, Mihai Paunescu and Lucy Walhain	132
The GAMETRAPP Project: Post-editing Neural Machine Translation of Research Abstracts in a Gamified Environment Cristina Toledo-Báez and Laura Noriega-Santiáñez	145
Google Translate Error Analysis for Mental Healthcare Information: Evaluating Accuracy, Comprehensibility, and Implications for Multilingual Healthcare Communication Jaleh Delfani, Constantin Orăsan, Hadeel Saadany, Özlem Temizöz, Eleanor Taylor-Stilgoe, Diptesh Kanojia, Sabine Braun and Barbara Schouten	151
Updating translator education programs: Adapting to technologies and their impacts in the Canadian language industry Elizabeth Marshman, Anwar Alftelawi, Haifa Ben Naji, Dipen Dave, Ahmed Elhuseiny Bedeir And Ting Liu	169
Studying the Need to Optimize Search for Amendments and Corrigenda in EU Institutional Translation Timea Palotai-Torzsas and Robin Palotai	186

Term Translation: Convert or Converse? Aida Kostikova, Kristin Migdisi, Sara Szoc and Tom Vanallemeersch	195
Hierarchical Data Linkage in a Terminology Management System: Challenges and Solutions at Bioleksipēdija Karina Šķirmante, Silga Sviķe, Arturs Stalažs, Gints Jasmonts and Roberts Ervīns Ziediņš	214
Exploring the integration of ChatGPT in academia and in the office: a preliminary case study Kyriaki Kourouni	221

New approaches to studying the cognitive impact of a CAI tool on Chinese interpreting trainees

Du Zhiqiang

University of Bologna

zhiqiang.du2@unibo.it

Ricardo Muñoz Martín

University of Bologna

ricardo.munoz@unibo.it

Abstract

High cognitive demands may impact (trainee) interpreters' performance when interpreting simultaneously, and computer-assisted interpreting (CAI) tools like InterpretBank seek to improve their term-rendering accuracy and efficiency. Part of a PhD research project, this paper reports on term rendering accuracy of Chinese interpreting trainees using CAI tools in remote interpreting tasks, and other cognitive effects. We adopted a control pretest-posttest design over three cycles. After an initial data-collection round (baseline), participants were split into either Excel or CAI-tool-trained groups for two more rounds, each with two tasks: glossary compilation and interpreting. We found noticeable improvements in term accuracy in both groups, but especially with the CAI tool, and complex interactions of CAI tool use with cognitive effort. InterpretBank, the CAI tool tested, seems to enhance term precision, decrease response times, and to support multitasking in remote interpreting, thereby improving interpreting quality. Yet results also reveal the cognitive challenges of information searching when the source speech delivers high-density information.

This paper reports on part of a PhD research project focused on the cognitive aspects of CAI-tool supported, simultaneous interpreting (SI) in remote environments. After reviewing the literature on CAI use (§ 1), the paper describes the research methods and dwells on some innovative details in data collection and constructs (§ 2). This is followed by data analysis (§ 3), and a summary of provisional results (§ 4).

1 Introduction

Computer-assisted interpreting (CAI) tools aim to enhance interpreters' work when extracting technical terms from documents, compiling glossaries, and retrieving terms from them (Fantinuoli 2018, p. 4). CAI tools such as InterpretBank (IB) have sparked considerable expectations—particularly, in *simultaneous interpreting* (SI)—due to their potential to support more accurate output and more efficient term management. In remote SI, CAI tools are expected to assist interpreters and ease continuous speech delivery by removing the need to search for pre-selected terms ~~previously~~ entered in a glossary beforehand.

The literature mainly focuses on term rendering accuracy, response time, and multitasking. Only accuracy is a typical indicator in studies of SI quality, particularly when SI is supported by CAI tools. Atabekova et al. (2018), Prandi (2018, 2020), and Defrancq & Fantinuoli (2021) have studied the use of IB and its reception among practitioners using European languages. In China, three masters' theses—Zhou (2019), Zhang (2021), and Ge (2023)—investigate its effects on trainee interpreters' performance and conclude that they improve term accuracy. Ge (2023) used a pretest-posttest design and reported a 23.1% improvement in term retrieval efficiency and 16.7% in term rendering accuracy in eight MA interpreting trainees, as judged

by four professional interpreters, who checked the matches of selected output terms in the glossary. These results suggest the benefits of using CAI tools and raise the question of whether to introduce it in SI training.

As for response times, time latency has become an important indicator for studying multilectal mediation tasks, and eyetracking and keylogging are popular tools for measuring it. In SI, *ear-voice span* (EVS) is deemed a useful measurement for assessing the effects of latency on interpreting quality. Timarová et al. (2011) found an average EVS of 4.1 seconds in 16 professional interpreters with Czech or Dutch as L1. Su (2020) compared the EVS of simultaneous interpreters working from English into Chinese, and found averages of between 0.93 and 3.25 seconds for novices and between 1.17 and 3.93 seconds for professionals, which suggests that experience may allow for a slight *increase* in time latencies.

Task conditions may also play a role and researchers of other tasks have drawn from EVS to develop parallel measurements. For instance, Chmiel & Lijewska (2022) examined *eye-voice span* (IVS)—the time latency in sight translation between output and eye movements—and found the average IVS in 24 conference interpreters to be over 8 seconds. Regarding CAI tool use, Fantinuoli & Montecchio (2023) contend that, when integrating automated suggestions, latency should not exceed the interpreters' average EVS as the limit for acceptable latency. They found that interpreters welcome automated suggestions and that, indeed, such prompts extend their EVS up to 2 seconds, i.e., just below their average EVS.

As for multitasking, as well as listening and talking at the same time, SI demands nearly flawless coordination of additional activities, e.g., with hands and eyes, such as interpreting visual signals from speakers, reading slide presentations, and managing consoles. This multitasking places high cognitive demands on interpreters, especially when (sub-)tasks share cognitive resources, potentially diminishing performance (Stachowiak 2014). In remote settings, an additional technical dimension becomes part of the SI workflow. In spite of the interest, progress in this area has been scarce. Human communication in digital environments can be studied as embedded in a broader, more complex process of human-computer interaction (HCI). From a cognitive-situated perspective, however, EVS only captures a single moment in time and does not adequately reflect the dynamic nature of tasks as they unfold or their full relationship with source speech delivery. Zhou et al. (2021) compared sentence-initial EVS with sentence-final EVS, yet this approach still falls short of providing a comprehensive view from an HCI perspective. On its own, EVS offers a limited snapshot of the broader cognitive processes, which are inherently parallel and involve multiple simultaneous activities. Relying exclusively on EVS risks overlooking the intricate interplay of tasks that extend beyond that specific moment.

By integrating current digital tools, we can enhance data quality and expand research possibilities. For example, glossary hits are typically assessed as correct or incorrect by matching the interpreter's recorded output with glossary entries, ignoring the fact that some hits may not be due to CAI tool support. Term repetitions highlight this issue, as repeated searches are treated as if the interpreter has no memory of previous glossary hits. Addressing these limitations will lead to a more accurate and nuanced understanding of interpreter behavior and cognitive processes in digital environments. As for expanding possibilities, keylogging makes it possible to cross-reference SI output with the activities of interpreters at the keyboard. Studying the quality, timeliness and success of searches and their use in the booth become possible. Eyetracking was not possible in this project, which collected data remotely.

This research project aimed to further investigate how interpreters use IB when confronted with technical terms. We were particularly interested in the timeline and interaction of interpreting subtasks when confronted with targeted terms. When using CAI tools, searches may coincide with the delivery phase, notably when the paired languages are distant and matching renderings may be more complicated, as in interpreting between English and Chinese.

This study focused on such instances, through features such as the length of source-speech segments. Longer segments compel informants to handle more information at once and this, in turn, may increase cognitive demands when using the CAI tool. Specifically, we studied those *overlaps* when the informants searched for terms within IB while they were presumably listening to the source speech. Given the variation in individual response times to problems, such overlaps capture the relative placement of actions in time regardless of whether they are considered swift serial activities or real multitasking, i.e., concurrent (sub-)task performance. This approach highlights the complex interplay between CAI tool use and the cognitive demands potentially placed on informants in SI.

To sum up, this study investigates the dynamics of using IB to support SI through ear-key span, search behaviors, and eye-voice span of Chinese interpreting trainees during term-intensive remote SI tasks. It seeks to shed light on how IB may impact their performance and to explore whether CAI tools might be a welcome addition in interpreter training programs. Thus, the research question is: How does IB influence the efficiency of term retrieval during remote interpreting sessions with term-dense source speeches? In other words, we want to determine whether IB improves Chinese interpreting trainees' handling of speeches rich in specialized terms.

2 Materials and methods

InterpretBank (IB) was chosen as representative of recent CAI tools. The feature of voice recognition and display of automatic retrieval was set to off to avoid additional variables that would have added further complexities to the design and perhaps mask some results.

2.1 Informants

Availability was an important criterion for the choice of informants, but not the main one. Trainees were preferred to professionals to foster comparisons with the three closest precedents, Zhou (2019), Zhang (2021), and Ge (2023). Furthermore, the trainees' lack of experience might foster higher frequencies of inefficient phenomena and breakdowns. Results in this project might lead to refining the methods, to better capture the potentially smaller effects in experts.

Professional interpreters often express concerns about participating in studies where they perceive their performance might be evaluated. (Englund Dimitrova et al., 2000). In contrast, students were easier to recruit and had more flexible schedules and thus were more available to participate in the study. Therefore, participant recruitment was conducted as convenience sampling among volunteering interpreting trainees. Informants were personally recruited by the researchers, but there was no prior acquaintance between anyone involved in the project.

The study involved 22 Chinese L1 and English L2 informants whose average age was 24.7 years (*s.d.* 2.9). All informants were trainees in MA's programs for conference interpreting at leading Chinese universities who had completed at least one year of SI training prior to this study. They were treated as a single group in the first data-collection round (baseline) and later

split into two groups, an *InterpretBank group* (IB, experimental) and an *Excel group* (XL, control, see § 2.3). The groups had 12 and 10 informants, respectively.

In cycle 3 (2nd post-test) the informants were allowed to use the tool of their choice. This resulted in one member in each group using a tool other than the one assigned to their group for cycle 2. To preserve the integrity of the data, these two informants were removed from the analysis of cycle 3, to ensure that the results accurately represented each group’s consistent and intentional use of either Excel or InterpretBank throughout the study.

The informants’ names have been replaced by nicknames in alphabetical order: A to L (Alex to Lee) belong to the IB group and M to V (Morgan to Val), to the XL group. Graphics and tables may show either the name or the initial, for reasons of space. In the text, full nicknames will be used.

2.2 Materials

Three texts on common health issues were selected from one English-language podcast series (Table 1). An L1 English professor, translator, and interpreter meticulously revised and vetted the shortened transcriptions of these texts to ensure their logical and natural flow and thematic unity. For consistency in interpretation standards, source speeches were recorded by three American interpreters who are L1 English speakers. Table 1 summarizes a few of the source speech recordings’ features. The acoustic properties, namely the number of syllables (*nsyll*), duration in seconds (*dur(s)*), and speech rate (*nsyll/dur*), were computed using the Praat software, following the script provided by de Jong et al. (2021).

speech	topic	word count	nsyll	dur(s) ^a	speech rate (nsyll/dur)
1	perception of time	1686	2558	776.35	3.29
2	immune system	1673	2383	793.75	3
3	emotions	1752	2470	777.53	3.18

Note. nsyll = number of syllables, dur = duration

^a start from the initial syllable to the final syllable

Table 1. Features of source speech recordings for booth tasks.

We introduced (*potential*) *problem triggers*—specialized terms and phrases of kinds known to cause interpretation difficulties—to focus the analysis. These triggers comprised 33 newly-presented terms. Furthermore, three of the terms were repeated twice in each text, so as to examine the effects of IB on the informants’ memories. This strategy aimed to mirror the complex conditions interpreters often face, to support empirical validity. All in all, and unknown to the informants, 39 technical terms were the targets to test the informants’ SI performance.

The study was conducted remotely due to COVID-19 restrictions. Participants needed a computer with headphones and a reliable Internet connection. They also installed a Python-based keylogging application and TechSmith screen recording software, both compatible with various operating systems, in their computers. These measures contributed to ensure a

controlled, noise-free, but quite natural task environment for high-quality data collection. IB was provided only after C1, and only to the IB group members, before they completed a training workshop on its use.

2.3 Methods

The study had a control pretest-posttest design, spanning three cycles (C1, C2 and C3), to analytically compare the effectiveness and user experience of InterpretBank and Excel in assisting term retrieval during SI tasks, and its potential progression through the cycles. Each cycle comprised a glossary-building task and an SI task, but here only the booth task will be addressed.

The primary independent variable was the use of technological aids for term retrieval during SI tasks. In C1, all informants used whichever information sources they preferred to build glossaries in Excel, a typical way to compile SI glossaries. Subsequently, the C1 data was used as a baseline of the performance of all informants. Based on the informants' experience with IB, they were divided into two cohorts: the group with no IB experience used it in C2 and C3, while those with some IB experience were assigned to the control group and continued with Excel in C2 and C3.

Data analysis was further developed for the IB group during C2 and C3 because an additional goal was to understand how the experimental informants' behavior evolved in a second post-test round (while helping to discern any impact derived from the novelty of using the tool, rather than from the tool use itself). Data for the XL group required the manual collection of data from screen recordings. The tasks in Excel often involved participants scrolling through entries or changing the targeted term during their search, which complicated the identification of precise moments associated with term retrieval due to factors such as screen size and width of visual field. This study aimed to favor ecological validity by not employing intrusive methods such as eyetracking devices that would have introduced additional variables or stress factors, potentially affecting informants' performance.

Procedures in C1, C2 and C3 were identical. Three distinct but similar speeches were interpreted across cycles (available upon request). In advance, the informants had compiled and sent their own glossaries on similar texts. They later received and adjusted a compiled master glossary 30 minutes before each SI task that included the relevant 33 problem triggers. Each glossary had ca. 100 terms altogether, with all lemmas that appeared at least in two individual glossaries.

The informants activated the screen recording and keylogging software before they began the SI task. Then they interpreted a single-play English speech recording and stopped interpreting at a designated end signal, 10 seconds after the source speech ended. C2 and C3 only differed in that about half of the informants used InterpretBank and the other half Excel. The IB group attended an online training workshop on the use of InterpretBank and engaged in self-directed trials to familiarize themselves with the use of the CAI tool during SI. They were allowed to choose when and how to employ it for term retrieval when encountering difficult terms in SI.

In this study, each informant accessed the source speech web player independently across all cycles, resulting in varying timelines in each screen recording and keylogging file. To make

comparisons possible, we synchronized all screen recording and keylogging timelines, plotting these events onto a universal timeline for each cycle.

2.4 Indicators

Performance was measured in various ways. For term accuracy, each rendering in a recording was classified as *correct* (term rendered with the glossary solution), *adequate* (good solution but not identical to the one in the glossary), *wrong*, or *dropped* (not rendered). Dropping source-text segments concerned only full clauses and sentences, since they are less likely to be the result of a strategic decision or choice by the interpreter. The indicators included eye-key span and eye-voice span. *Ear-key span* is defined as the latency between the end of the soundwave of the source speech utterance of a potentially problematic terminological unit (often plurilexical) to keydown of the informant's first related keyboard action. *Eye-voice span* (cf. Chmiel & Lijewska 2022) in our study measured the latency between onscreen display of a glossary entry and the start of the soundwave by the informant vocalizing the corresponding Chinese voice rendition. Both indicators could only be measured in the IB group—this was not possible with the Excel setup. Additionally, we measured the duration of search activities and the length in milliseconds of dropped source speech segments, as uniformly chunked beforehand by the researchers (most of them sentences).

3 Results and discussion

The analysis is grounded on aligned keylogging and screen recording data, and aims to explore the influence of using a CAI tool (InterpretBank) on the informants' SI output and cognitive processes during the SI tasks across cycles.

3.1 Term accuracy

The 33 targeted potential problem triggers appearing in the source speech *for the first time* are referred to here as *normal terms*. Figure 1 displays the distribution of rendering quality categories in percentages. The IB group enjoyed a significant improvement over the cycles. In C1, only about 19.8% of the terms were correctly interpreted, establishing a relatively low baseline for the group. In C2, the percentage of correctly interpreted terms increased to approximately 34.7%—almost 15% increase from C1 in absolute terms, and 57% in relative terms). There was also a rms, so that they can be said to have displayed an overall improvement in their performance. The IB group performed even better in C3, with approximately 50.4% of the terms interpreted accurately, reinforcing the upward trend observed between C1 and C2.

The XL group, in contrast, also showed steady improvement although it was more modest. In C1, they *correctly* interpreted 22.2% of the terms. In C2, the figure increased to slightly more than one out of four (28.3%). In C3, the XL group managed to *correctly* interpret around 37.7% of the terms, or about two out of five, revealing a comparatively smaller but steady improvement. In brief, both groups improved between C1 and C3, but the IB group showed a more significant increase in correctly interpreted terms. The data thus suggests that InterpretBank may be an effective resource in helping interpreters achieve higher accuracy with unfamiliar terms during interpretation.

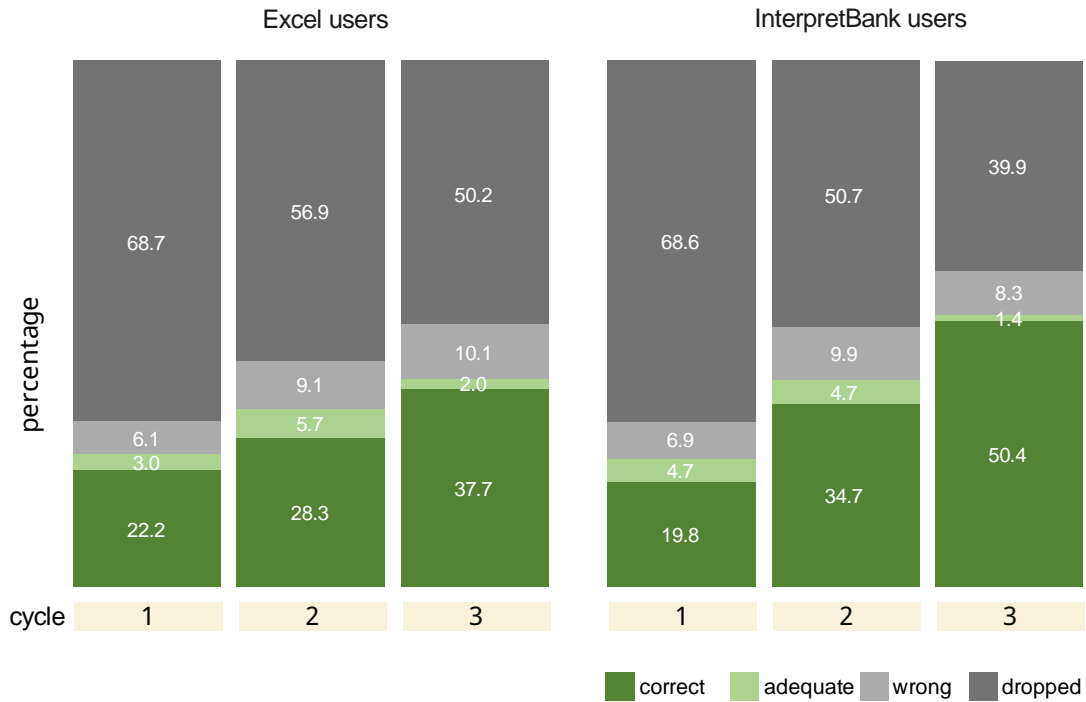


Figure 1. Stacked percentage bar plot of *normal terms* by cycle and group.

Let us now turn to repetitions. Three of the normal terms discussed above were intentionally repeated twice throughout each source speech. The first instance is labeled *normal+rep* in Table 2, to remind the reader that they were also part of the 33 terms discussed in Table 1. *Rep1* refers to the first repetition of the term, i.e., when it appeared for the second time in the source speech. These terms are repeated once after their initial occurrence. *Rep2* refers to the second repetition or third time the term was uttered.

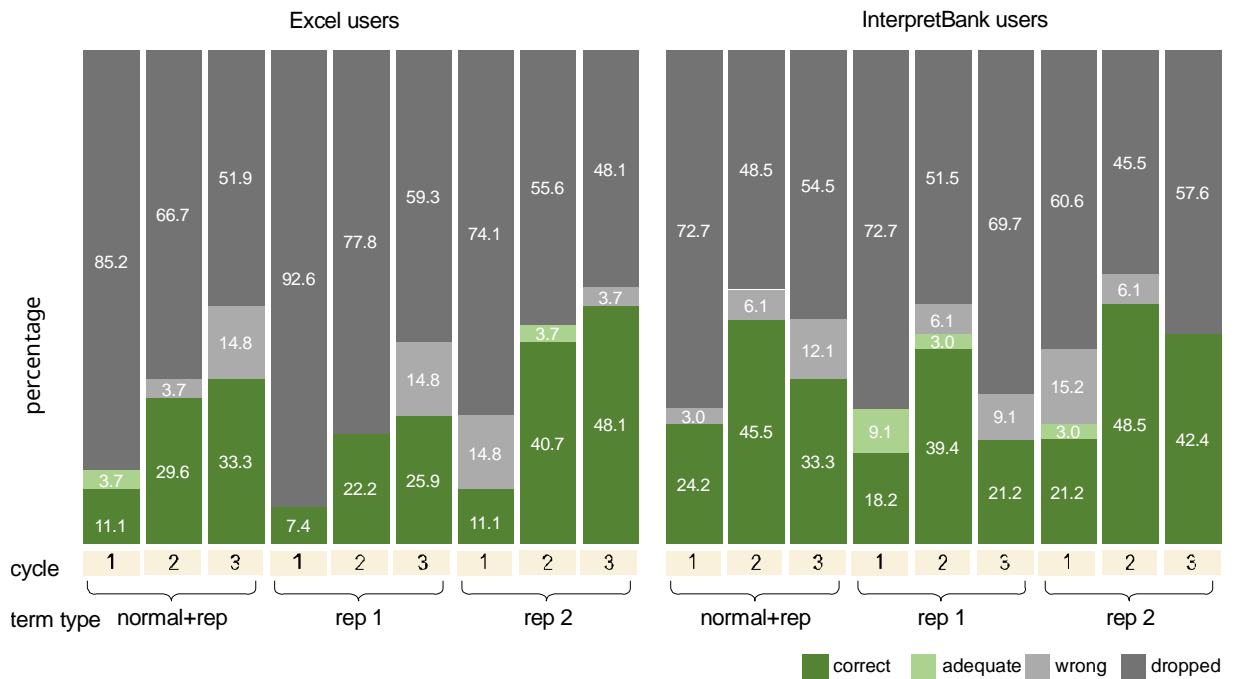


Figure 2. Stacked percentage bar plot of *repetition* terms by cycle and group.

Figure 2 shows that the XL group experienced improvements when interpreting the *normal+rep* terms, confirming that their behavior was similar to that with the rest of the terms appearing for the first time. The figures are different, but the small sample does not allow us to draw any further conclusions beyond the upward tendency. Starting at a modest 11.1% rate in C1, the XL group improved to 29.6% in C2 and reached 33.3% in C3. In *rep1*, the group experienced a sharp rise from a low 7.4% rate in C1 to 22.2% in C2, before tempering down or leveling off at 25.9% in C3. This suggests a likely adaptation to the task during C1 or else a learning curve that gained momentum in C2, whose improvements either diminished or started to plateau. For *rep2*, the group made a dramatic improvement from C1's 11.1% rate to C2's 40.7%, to their peak of 48.1% correct term renderings in C3. This suggests a sustained progress across the cycles. More importantly, repeated terms might have a different impact than when repeated for the first and second times (second and third appearances).

The IB group's percentage of correct interpretations in the *normal+rep* (first time) displayed a bell curve. The group began at 24.24% in C1, peaked at 45.5% in C2, and then fell to 33.3% in C3. Interestingly, the same happened in C2 and C3. The *rep1* category experienced a peak in C2 with 39.4% correct interpretations, a significant leap from C1's 18.2%. However, the group's performance then experienced a decline, settling at 21.2% in C3. In *rep2*, the IB group displayed an upward trend that peaked at 48.5% in C2, and more or less plateaued at a strong performance of 42.4% in C3. This might be interpreted as effective learning or adaptation across the cycles as well.

The performance patterns across the three cycles were thus different between the groups. The correct output from the XL group generally yielded a progressive increase (*normal terms*) while the IB group showed higher improvements overall. In any case, the combined data displayed in Figures 1 and 2 suggests that both groups improved in all types of terms across the cycles.

However, with repeated terms (i.e., *normal+rep*, *rep1*, *rep2*) the performance of IB group peaked at C2 but experienced a decline in C3. Initially, we would expect the informants to take the first repetition (second appearance) of a term as a strong hint to keep it active in working memory for future needs. That is, our expectation was exactly the opposite, that the IB group informants would tend to use InterpretBank's prompts less while rendering more terms well. A possible explanation is that the informants might have remembered the glossary equivalent at *rep1* but, now confident that they could rely on the glossary, they might have wiped the term out of their working memories as they went on—or, rather, let it naturally fade away, reassured as they were that they could rely on InterpretBank. Their attention might have shifted elsewhere, as they presumed that the term was no longer problematic.

In contrast, the XL group might have relied more on term recognition—that is, on keeping repeated terms active in memory. This analysis hints at a potentially nuanced impact of tool use on cognitive engagement with repetitive terms, and how the choice of tool could influence memory retention and attention allocation during the interpreting task. However, the data does not allow us to discern whether the IB informants actually used InterpretBank's prompted output or just proceeded without it. They might have felt prompted to continually type the words, thereby enhancing their memory retention of repetitive terms. The bottom line is that this scenario suggests that using InterpretBank does not necessarily guarantee that repeated

source speech terms will be rendered correctly, indicating an area that deserves further research.

3.2 Ear-key span and eye-voice span

Ear-key span measurements shed light on the informants’ auditory processing speed, comprehension, and decision-making in terms of whether to resort to a CAI tool, such as InterpretBank, for assistance. A shorter ear-key span is interpreted to indicate faster comprehension or decision-making processes, whereas a longer span may suggest challenges in understanding or a slow realization that some kind of support is necessary.

Eye-voice span is hypothesized to measure the informants’ visual processing speed and their ability to synthesize and integrate the displayed translation suggestions into their outputs. A shorter eye-voice span is taken to suggest swift assimilation of visual rendering suggestions from IB or rapid articulation into the target speech, while a longer span could hint at challenges in synthesizing the information retrieved or in fluently delivering it.

	<i>Cycle</i>	<i>Mean</i>	<i>Median</i>	<i>Mode</i>	<i>SD^b</i>	<i>Minimum</i>	<i>Maximum</i>	Shapiro-Wilk (not ms)	
								<i>W</i>	<i>p</i>
ear-key span	2	1925	1805	1078 ^a	1392	-1498	8701	0.948	< 0.05
	3	1639	1083	-0797 ^a	2227	-1272	8858	0.865	< 0.05
eye-voice span	2	2309	2019	1214 ^a	1785	-2600	8732	0.954	< 0.05
	3	1503	1689	1515 ^a	2299	-7778	7473	0.821	< 0.05

Note. ^a More than one mode exists, only the first one is reported.

^b SD = standard deviation

Table 2. Descriptive statistics and Shapiro-Wilk test results of ear-key spans and eye-voice spans, in milliseconds.

The mean ear-key span value for C2 was 1925 ms (Table 2)—slightly higher than the median of 1805 ms. This suggests a fairly regular data distribution, but the standard deviation of 1392 ms indicates moderate variations within the data. The range of values stretched from a minimum of -1498 ms to a maximum of 8701. The Shapiro-Wilk test for normality indicated that the data are not normally distributed. For C3, the mean ear-key span dropped to 1639 ms, with a median of 1083 ms, and the main (most usual) mode was reported at -797 ms, although multiple modes exist. The standard deviation increased to 2227 ms, showing greater variations among the informants. The Shapiro-Wilk test resulted in a *W*-value of 0.865 ($p < 0.05$), which confirms a non-normal distribution.

As for the eye-voice span, in C2 the mean was 2309 ms; the median, 2019 ms; and the main mode, 1214 ms, with the caveat that multiple modes exist. The standard deviation was 1785 ms, and the values ranged from -2600 ms to 8732 ms. The Shapiro-Wilk test ($W = 0.954$, $p < 0.05$) strongly suggested a non-normal distribution. In C3, the mean eye-voice span was lower

at 1503 ms, with a median of 1689 ms and a mode of 1515 ms. The standard deviation here was 2299 ms, i.e., there was a broader spread of data, ranging from -7778 ms to 7473 ms. The Shapiro-Wilk test for this cycle pointed to a W -value of 0.821 with p lower than the conventional alpha of 0.05, so the data were not normally distributed. Overall, the statistical metrics confirmed non-normal distributions and different degrees of variation for both ear-key span and eye-voice span across the two cycles. The means and the medians suggest a shift in central tendencies between the cycles, and the significant p -values from the Shapiro-Wilk tests underscore the need for non-parametric analyses (e.g., Kendall's Tau-b below).

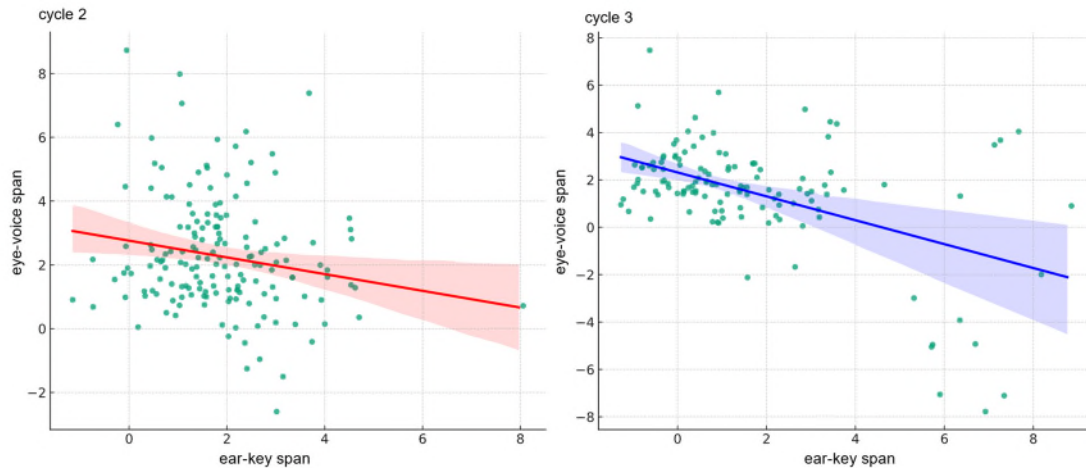


Figure 3. Scatter plots of ear-key and eye-voice span across C2 and C3.

The scatter plots in Figure 3 offer complementary perspectives on the relationship between ear-key span and eye-voice span across Cycles 2 and 3. C2 displayed a wide distribution of data points. The broad 95% confidence interval suggested a high degree of variation and uncertainty. To explore the correlation between ear-key span and eye-voice span, we conducted a Kendall's Tau-b statistical test, which yielded a value of -0.103 , $p = 0.048$. These figures point to a weak but statistically significant negative correlation between the two variables. Additionally, the wide 95% confidence interval on the scatter plot signaled a considerable level of uncertainty. The right scatter plot for C3 also showed a widely dispersed range of data points, and the 95% confidence interval corroborated the high level of variation observed in C2. Kendall's Tau-b was -0.228 , $p < 0.01$, which implies a weak to moderate but statistically significant negative correlation between the two variables. That is, the longer they took to start searching for a term, the shorter they needed to then integrate the IB's prompt into their output.

While the scatter plots for both C2 and C3 suggest weak correlations, the Kendall's Tau-b values and p -values provide a clearer picture: they confirm a statistically significant negative correlation, weaker in C2 than in C3, though still weak to moderate in the second case. Thus, the decrease in mean ear-key span between C2 and C3 suggests that informants might have become more efficient at processing auditory cues or more resolute in deciding when to use IB. This could be the result of increased familiarity with the experimental task procedures or enhanced proficiency with the tool.

3.3 Search duration and dropped chunks

Recordings from the IB group in both C2 and C3 were aligned on their respective universal timelines. In C2, the common timeline spanned from 0 to 803.804 seconds. In C3, the common timeline extended from 0 to 788.577 seconds.

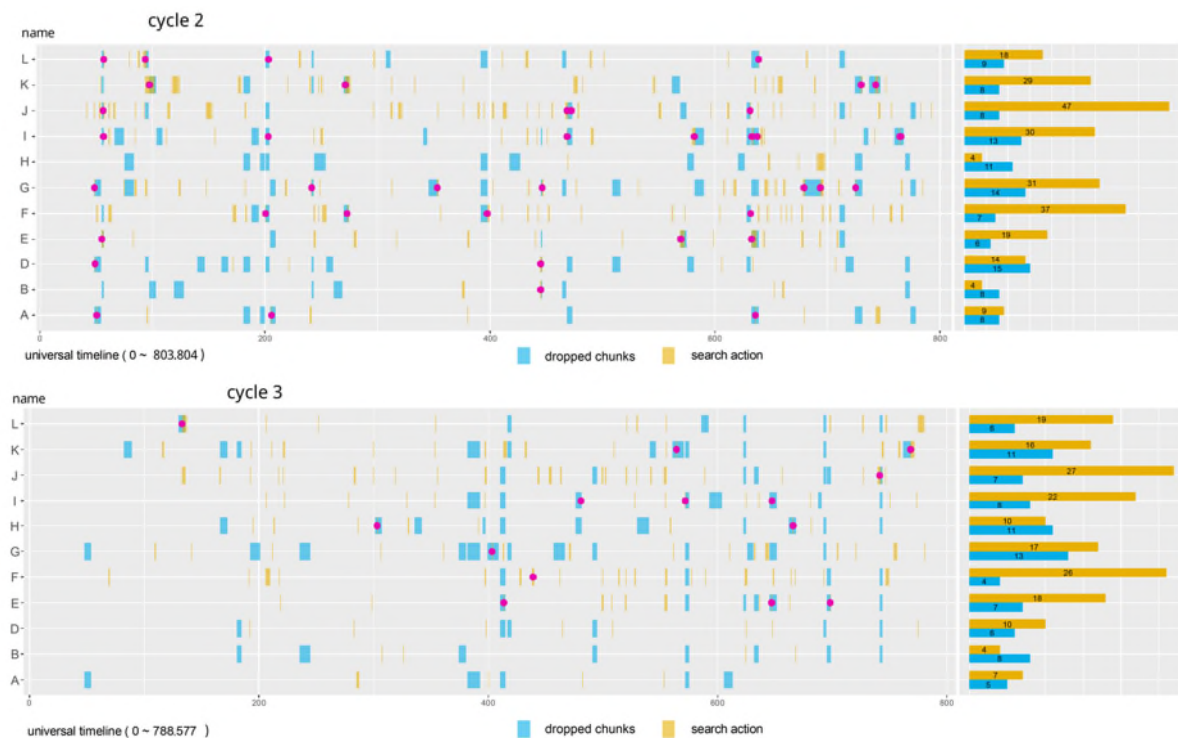


Figure 4. Individual distribution of search events and durations with overlaps in Cycles 2 and 3.

In Figure 4, the yellow blocks represent the time spans (duration) of each search in shared (aligned) timelines, while the blue areas denote chunks of the source speech that were dropped. Overlaps between searches and dropped source-speech fragments are marked by pink dots and are interpreted as failed attempts to interpret the sentence. that failed

In Figure 4, the upper diagram represents the second cycle. C2 exhibited a distinctive distribution of overlaps (pink dots). Notably, the session began with 14 overlaps, then 10 in the middle and, intriguingly, showed a slight uptick with 16 overlaps towards the end. On the other hand, C3—with a timeline close to 789 seconds—told a different story. Overlaps were notably sparse at the beginning of the SI task, with just a single occurrence; this increased moderately to 5 in the middle, and escalated to 8 towards the end. The combined figures show that overlaps were not uniformly distributed across the interpreting sessions. While C2 presented a fairly balanced distribution with a slight skew towards the end, C3 markedly shifted this balance, the end phase of the task now critical.

The informants also presented unique patterns in their individual performances. For instance, Alex had eight *dropped chunks* and nine *searches*, with three overlaps occurring at 50.43, 205.86, and 635.98 seconds. Erin, on the other hand, exhibited seven *dropped chunks* and eighteen *searches*, and also experienced three overlaps at 413.49, 647.10, and 698.23 seconds. These individual patterns varied not only between *dropped chunks* and *searches* but also in the frequency and timing of their overlaps. We cannot here discuss these interesting results in further detail.

A particular trend emerged across both cycles, where searches often closely corresponded to dropped chunks, especially in C2. This suggests that term retrieval, more or less facilitated depending on the tool they used, may have played a role, alongside other cognitive factors. These observations show that searches and dropped chunks frequently occurred

simultaneously. Their co-occurrence probably reflects the multitasking demands placed on the informants when performing two subtasks at once, namely processing the source speech while searching for terms. Of course, the impact of this multitasking might vary depending on the type of speech, and its impact on the informants is very likely related to factors such as experience and expertise. But the sample interpreted the same speeches in the same order and all other testing conditions were very similar, if not identical, so that tentative explanations for variation might rather link to aspects such as working memory, processing speed, strategies to cope with stress, and even personality.

4 Conclusions

A control pretest and posttest study examined the performance of interpreting trainees as they engaged in remote simultaneous interpreting with different CAI tools (Excel and InterpretBank) in three task cycles—baseline (pre-test) and two post-test rounds—under otherwise identical conditions. The results as regards term accuracy revealed that, compared to the Excel group (XL), the InterpretBank (IB) group showed higher improvements in the correct rendering of terms they faced for the first time. However, repeated terms revealed a different pattern: the XL group tended to improve through the three cycles whereas the IB group bounced back to a more modest performance from cycle 2 to cycle 3.

To the best of our knowledge, this study was the first one to incorporate keylogging into interpreting studies. Thanks to the time-stamps of keystroke and mouse events, we used ear-key span and eye-voice span as indicators, which decreased through the cycles in both groups and showed a weak negative correlation with each other. This trend suggests that informants become increasingly apt at quickly identifying problem triggers, and at making rapid decisions as to whether they should seek support to render them.

Behavioral patterns also suggest that informants may rely on InterpretBank when they feel challenged by cognitive demands. Individual behavior revealed substantial variations in terms of the number and duration of search events across cycles. These variations were closely aligned with the unfolding of the speech. Specifically, in C2, searches increased in number at certain spots at the beginning (25%), middle (50%), and end (25%) phases. In contrast, in C3, searches concentrated on spots predominantly in the middle and towards the end of the speech. We were very careful to avoid differences between texts, but an impact cannot be ruled out. This variation highlights the dynamic nature of the interactions between informants and InterpretBank.

This report is necessarily limited, but it clearly suggests that this research question deserves further study. A situated perspective on CAI-tool supported remote SI as an instance of human-computer interaction permits a more complete and rigorous analysis with increased data accuracy. Combining keylogging with screen and sound recording captures relevant aspects of SI multitasking and fosters new hypotheses and research paths.

Acknowledgements

We would like to express our gratitude to all the informants who participated in this study.

References

Atabekova, A. A., Gorbatenko, R. G., Shoustikova, T. V., & Valero-Garcés, C. (2018). Cross-cultural mediation with refugees in emergency settings: ICT use by language service

- providers. *Journal of Social Studies Education Research*, 9(3), Article 3. <https://jsser.org/index.php/jsser/article/view/274>
- Chmiel, A., & Lijewska, A. (2022). Reading patterns, reformulation and eye-voice span (IVS) in sight translation. *Translation and Interpreting Studies*. <https://doi.org/10.1075/tis.21021.chm>
- de Jong, N. H., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, 28(4), 456–476. <https://doi.org/10.1080/0969594X.2021.1951162>
- Defrancq, B., & Fantinuoli, C. (2021). Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target. International Journal of Translation Studies*, 33(1), 73–102. <https://doi.org/10.1075/target.19166.def>
- Fantinuoli, C. (2018). Interpreting and technology: The upcoming technological turn. In C. Fantinuoli (Ed.), *Interpreting and technology* (pages 1–12). Language Science Press. <https://doi.org/10.5281/ZENODO.1493289>
- Fantinuoli, C., & Montecchio, M. (2023). Defining maximum acceptable latency of AI-enhanced CAI tools. In Ó. Ferreiro-Vázquez, A. T. Varajão Moutinho Pereira, & S. L. Gonçalves Araújo (Eds.), *Technological Innovation Put to the Service of Language Learning, Translation and Interpreting: Insights from Academic and Professional Contexts* (pages 213–225). Peter Lang Verlag. <https://doi.org/10.3726/b20168>
- Ge, T. (2023). *Usability of Terminology—Assistance in Chinese to English Simultaneous Interpretation—Taking interpreterBantias an Example* [Master's Thesis, Beijing Foreign Studies University]. <https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFDTEMP&filename=1023063456.nh&v=> [last accessed October 20, 2023].
- Prandi, B. (2018). An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation. In C. Fantinuoli (Ed.), *Interpreting and technology* (pages 28–59). Language Science Press. <https://doi.org/5281/ZENODO.1493281>
- Prandi, B. (2020). The use of CAI tools in interpreter training: Where are we now and where do we go from here? *inTRAlinea, Special Issue: Technology in Interpreter Education and Practice*. <http://www.intralinea.org/specials/article/2512>
- Stachowiak, K. (2014). Mind's not lazy: On multitasking in interpreters and translators. *Konińskie Studia Językowe*, 2(3), 293–313.
- Su, W. (2020). *Eye-Tracking Processes and Styles in Sight Translation*. Springer Singapore. <https://doi.org/10.1007/978-981-15-5675-3>
- Timarová, S., Dragsted, B., & Gorm Hansen, I. (2011). Time lag in translation and interpreting: A methodological exploration. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Benjamins Translation Library* (Vol. 94, pages 121–146). John Benjamins Publishing Company. <https://doi.org/10.1075/btl.94.10tim>
- Zhang, J. (2021). *An Experiment Report on the Impact of Computer-Aided Interpreting Tools on Simultaneous Interpreting* [Master's Thesis, China Foreign Affairs University].

<https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202201&filename=1021596437.nh&v=> [last accessed October 20, 2023].

Zhou, H., Weng, Y., & Zheng, B. (2021). Temporal Eye-Voice Span as a Dynamic Indicator for Cognitive Effort During Speech Processing: A Comparative Study of Reading Aloud and Sight Translation. In R. Muñoz Martín, S. Sun, & D. Li (Eds.), *Advances in Cognitive Translation Studies* (pp. 161–179). Springer Singapore. https://doi.org/10.1007/978-981-16-2070-6_8

Zhou, L. (2019). *The Impact of Computer-Aided Interpreting Tools on Simultaneous Interpreting Performance: Taking InterpretBank as an Example* [Master's Thesis, Xiamen University].

<https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202002&filename=1019069326.nh&v=> [last accessed October 20, 2023].

Poster on ChatGPT Translator Plus

Terence Lewis

Translator & Software Developer
support@mydutchpal.com

Abstract

This short paper describes the ChatGPT Translator Plus application developed to facilitate use of the translation capabilities of ChatGPT/GPT-4

This short paper describes ChatGPT Translator Plus. This application provides a desktop gateway to OpenAI's principal large language models (LLMs), namely ChatGPT-3.5-Turbo and GPT-4. The app is written in Python with a user-friendly GUI with a row of buttons extending beneath an input and output screen. These buttons offer the user a series of actions which are performed either locally or remotely by the selected model.

The primary purpose of this application is to facilitate the use of the translation capabilities of OpenAI's large language models.

Users are required to have an OpenAI API key which is obtained from the OpenAI website. API keys are available free of charge for a trial period, after which a subscription is needed. Once the API key is entered the main GUI shown below is displayed. It is only necessary to enter the API key once.

ChatGPT Translator Plus is essentially a provider of machine translations, but there are also buttons for the related tasks of Paraphrasing (Rewrite), Correction / Proofreading and Summarization. While any combination of source and target languages from amongst the *Flores 200* languages displayed in the combo boxes can be selected (and a translation will be attempted), ChatGPT-3.5-Turbo and GPT-4 have been found to perform best when translating high-resource languages into English.

The "Model" button enables users to select from GPT-3.5-turbo, GPT4 and later models. Models not included in the list box can be entered manually.

The "Temperature" button is used to set the temperature, which is a parameter controlling how random a large language model's output is likely to be. With a higher temperature, the model takes more risks, producing a translation that is creative and less predictable.

The "Chunk size" button plays a key role in how the model translates, and particularly how it can be used to translate large documents. OpenAI's models have a context window, a term used to denote the number of tokens that can be passed to and from the model in the combined prompt and response. GPT-4-Turbo currently has a context window of 128K tokens. Using the "chunk size" feature we can break a very large document down into the most appropriate size for a particular translation task.

The "Apply examples" feature allows the user to provide a file containing example translations of terms or phrases to be used by the AI model during the translation process. ChatGPT can handle quite complex examples that go beyond mere lists of words and phrases, so this feature enables the user to provide detailed instructions on the translation of technical

terms. Since the information is being provided from outside the LLM, this feature may be regarded as a simple form of Retrieval Augmented Generation (RAG).

LLMs are capable of performing a broad range of language-related tasks. This app focuses on translation and a few associated tasks.

The "Import document" button allows the user to import a docx, pdf or text file of any length into the input window. The document can then be translated by clicking the "Translate screen" button, the text being broken down into chunks via the "Chunk size" button. This is one way to do "document level" translation via ChatGPT/GPT-4.

The "Translate file" button provides the translation of Word, Excel, PowerPoint, text or PDF files, sending source text to ChatGPT and receiving the translation sentence by sentence. This model has less context knowledge than when the "Import document" feature is used but source and target file correspond line by line and the app also generates a TMX file enabling import into a standard CAT tool like Trados or MemoQ.

Successful use of ChatGPT Translator Plus for complex translation tasks is dependent on the creation of a well-targeted prompt. Our Prompt Manager offers various ways of designing a successful prompt, which include retrieval of an effective prompt from a prompt library and requesting the GenAI model to generate its own prompt. All prompt suggestions can be edited and adapted before being applied to the request sent to the GenAI model. A typical prompt could refer to the desired register and tone of the target text and the intended readership and request the model to apply a terminology list or to take into account various authoritative reference works.

The "Domain" button enables the user to specify the domain or specialist field covered by the document to be translated. This domain reference will be combined with the output of the Prompt Manager in the request sent to the GenAI model.

The "Save output" button allows the user to save output from the model directly to a disk. Use of this button is not necessary when the "Translate file" feature is used as the translation is then automatically written to a file on disk.

This description reflects the state of the application on 18/01/2024. ChatGPT Translator Plus is under continuous development and new features are likely to be added every few months.

Since the presentation and the submission of this short paper, new features have been added to the application. The most important of these is the provision of a Prompt Management module which enables users to store a library of prompts, edit these previously used prompts, write new prompts and even ask the model to suggest prompts for a particular task. This module has effectively replaced the "Apply examples" function since it enables the user to provide lengthy documentation as a guidance for the model.

Single-user licences for the software may be purchased from the author's website at <https://mydutchpal.com/shop>. Multi-seat licences can be purchased by contacting support@mydutchpal.com.

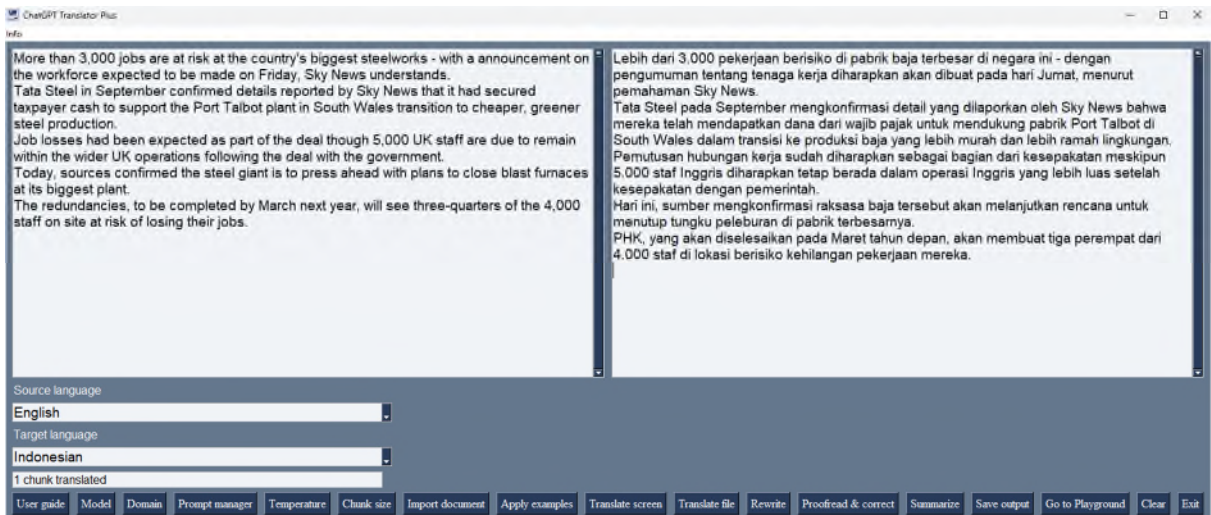


Figure 1. ChatGPT Translator Plus in translation mode

Large Language Models – origin, growth and capabilities

Andrzej Zydrón

XTM International

azydron@xtm.cloud

Rafal Jaworski

XTM International

rjaworski@xtm.cloud

Szymon Kaczmarek

XTM International

skaczmarek@xtm.cloud

Abstract

The rapid evolution of Large Language Models (LLMs) marks a significant milestone in the latest developments within artificial intelligence. These models characterized by their vast number of parameters and the corresponding breadth of language understanding have transformed natural language processing and generated a burgeoning field of study. This paper provides a comprehensive overview of the technical advancements that have catalyzed the growth of LLMs with a particular focus on the scalability of model parameters and the expansion of context windows. It delves into the emergent discipline of prompt engineering which includes techniques ranging from basic instruction crafting to sophisticated methods like role-play and Chain of Thought (CoT) prompting. Each technique is explored in depth emphasizing its utility and the nuances of its application. The paper also addresses the limitations of LLMs, especially in few-shot and zero-shot learning paradigms and the challenges these models face including biases and ethical considerations.

1 Introduction

The introduction of Large Language Models (LLMs) has significantly changed the field of artificial intelligence, especially in how computers understand and generate human language. At the heart of this transformation lies the transformer architecture, a groundbreaking model introduced by (Vaswani et al. 2017) which has since become the backbone of most state-of-the-art language processing systems. The transformer's unique ability to process sequences of data in parallel, leveraging self-attention mechanisms, has unlocked unprecedented capabilities in generating, understanding and interpreting human language at scale.

The rapid growth and evolution can be attributed to several key advancements. Firstly, the scaling of parameters has demonstrated a seemingly straightforward yet profoundly effective approach to improving performance; models with billions of parameters have shown remarkable abilities in understanding and generating complex texts. This parameter scaling, however, is not without its challenges including, but not limited to, computational efficiency and environmental impacts (AI Index Steering Committee, 2023; Luccioni et al., 2023).

Furthermore, broadening the context windows of models has significantly improved their understanding of longer texts enabling more coherent and contextually accurate outputs. These technical advancements have paved the way for innovative applications of LLMs, particularly in the realm of prompt engineering. Prompt engineering has emerged as a crucial technique for effectively interacting with LLMs, enabling users to guide the models' outputs through carefully designed inputs. This approach has unlocked many applications, from creative writing assistance to complex problem-solving tasks.

Despite their impressive capabilities, LLMs face several limitations and challenges, including issues related to hallucinations (Huang et al., 2023; Tonmoy et al., 2024; Agrawal et al., 2023; Zhang et al., 2023; Cheng et al., 2023; Zhang et al., 2023; Guerreiro et al., 2023), biases (Navigli et al., 2023; Abid et al., 2021; Gallegos et al., 2023; Ferrara, 2023), privacy concerns (Kshetri, 2023; Li et al., 2023), and potential misuse (Pan et al., 2023; Zhong and

Wang, 2023). Addressing these challenges is crucial for the responsible development and deployment of these technologies.

This paper aims to explore the technical advancements in LLMs, delve into the nuances of prompt engineering, and discuss the limitations and challenges these models face. By examining these aspects, we seek to provide a comprehensive overview of the current state and prospects of LLMs in the field of artificial intelligence.

2 Technical advancements in LLMs

The rapid evolution of the LLMs can be primarily attributed to the significant advancements in several key areas. These improvements have not only enhanced the capabilities of LLMs but have also expanded their application across various fields. Below, we explore the pivotal advancements that have marked the evolution of LLMs.

2.1 Parameter scaling and efficiency

A visual inspection of the trends in the development of generative models since 2019 reveals a noteworthy trajectory in parameter scaling. The graph represented in Figure 1. illustrates the progression of various models, with a clear upward trend in the number of parameters, culminating in models like PaLM with an astonishing 530 billion parameters (Chowdhery et al., 2022). The trajectory initially supported the hypothesis that the larger models would yield better performance.

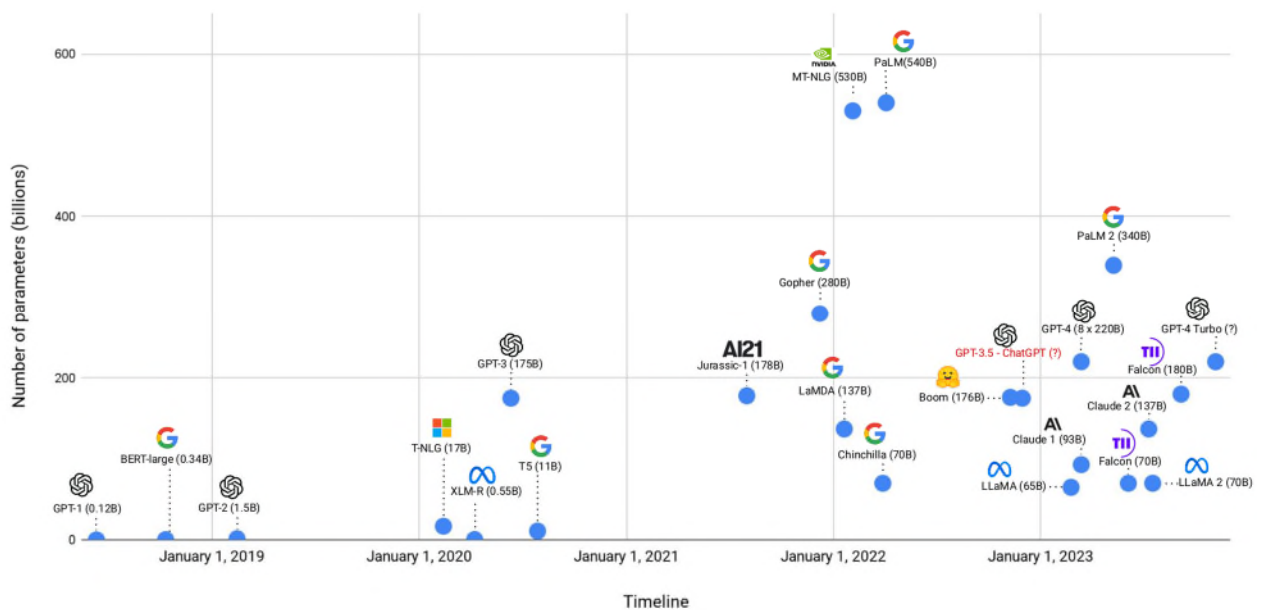


Figure 1. Evolution of generative models since 2019, plotted against the number of parameters

However, the release of the Chinchilla model marked a pivotal moment in the understanding of scaling laws for LLM (Hoffmann, 2022). Contrary to the previous trend, Chinchilla demonstrated that optimizing the ratio between parameters and training dataset tokens could yield performance similar to older and significantly larger models. This revelation was further reinforced by the development of the LLaMA (Touvron, 2023a) model, which also features fewer parameters than the largest models, but benefits from a much larger training dataset. The

result is a model that maintains high performance while mitigating some of the computational and environmental costs associated with very large models.

To make the relationship between model size and dataset scale clearer, Table 1 offers a comparison of major LLMs such as GPT-3, PaLM, LLaMA, LLaMA 2 (Touvron, 2023b), PaLM 2 (Anil et al., 2023), and Falcon (Almazrouei et al., 2023). This comparison focuses on three important aspects: the number of parameters, the size of the dataset in tokens, and the ratio between the two.

Model	Number parameters	of Dataset size (tokens)	Dataset size to parameters ratio
GPT-3	178B	499B	2.8
PaLM	540B	780B	1.4
LLaMA	65B	1.4T	21.5
LLaMA 2	70B	2T	28.5
PaLM 2	~340B	3.6T	10.5
Falcon	180B	3.5T	19.4

Table 1. Comparative Analysis of LLMs by Parameters and Dataset Size

The transition to a higher dataset-parameters ratio, as seen in models like LLaMA 2, suggests a refinement in the approach to training LLMs. Based on the benchmarks (Touvron, 2023a; Touvron, 2023b), models trained on more extensive and diverse datasets achieve higher accuracy without the need for an exponential parameter increase. This strategic scaling suggests a departure from the premise that more parameters equate to better performance. Instead, it highlights the importance of the quality and size of the training dataset.

2.2 Context window enhancements

Improvements in the amount of context that Large Language Models can consider have played a key role in improving their effectiveness. The context window is the amount of text – measured in tokens – that the model can actively consider during text generation. This window is crucial for the model’s understanding of text and is a key factor in the coherence and continuity of the generated text. Table 2 illustrates the expansion of the context window in LLMs over time. Early models like GPT-3 had a relatively modest context window of 2,000 tokens (Brown et al., 2020), but more recent developments have seen a substantial increase, with models like GPT-4 Turbo boasting a context window of 128,000 tokens (OpenAI, 2023).

Model	Number of parameters
GPT-3	2k
GPT-3.5 Turbo	16k
LLaMA	32k
GPT-4	32k
Claude 2	100k
GPT-4 Turbo	128k

Table 2. Context window sizes in recent LLMs

The increasing size of the context window suggests an improvement in the models' ability to process and generate text based on a broader context. This enables more detailed responses, particularly in tasks that require referencing information from earlier in text or conversation.

However, a larger context window does not necessarily equate to an ability to effectively utilize the entire span of tokens (Liu et al., 2023). Benchmarks such as "Needle In A Haystack - Pressure Testing LLMs" (Kamradt, 2023) reveal that models like Claude 2.1 (Anthropic, 2023), despite having a larger context window, exhibit challenges in retrieving accurate information from longer documents. The tests conducted as a part of this benchmark indicate that retrieval accuracy drops significantly as the document length increases. The reason for that is the fact that LLMs are proven to work better when presented with more specific information, e.g. in few-shot learning scenarios. The findings of this benchmark are illustrated in Figure 2 and Figure 3.

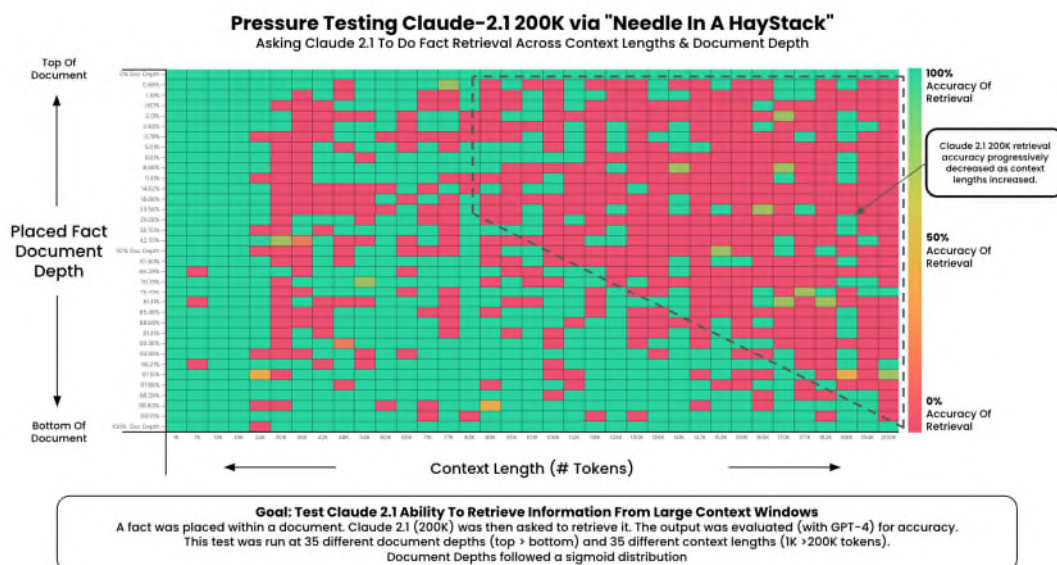


Figure 2. Pressure testing of GPT-4's context window via the "Needle In A Haystack" benchmark (Kamradt, 2023)

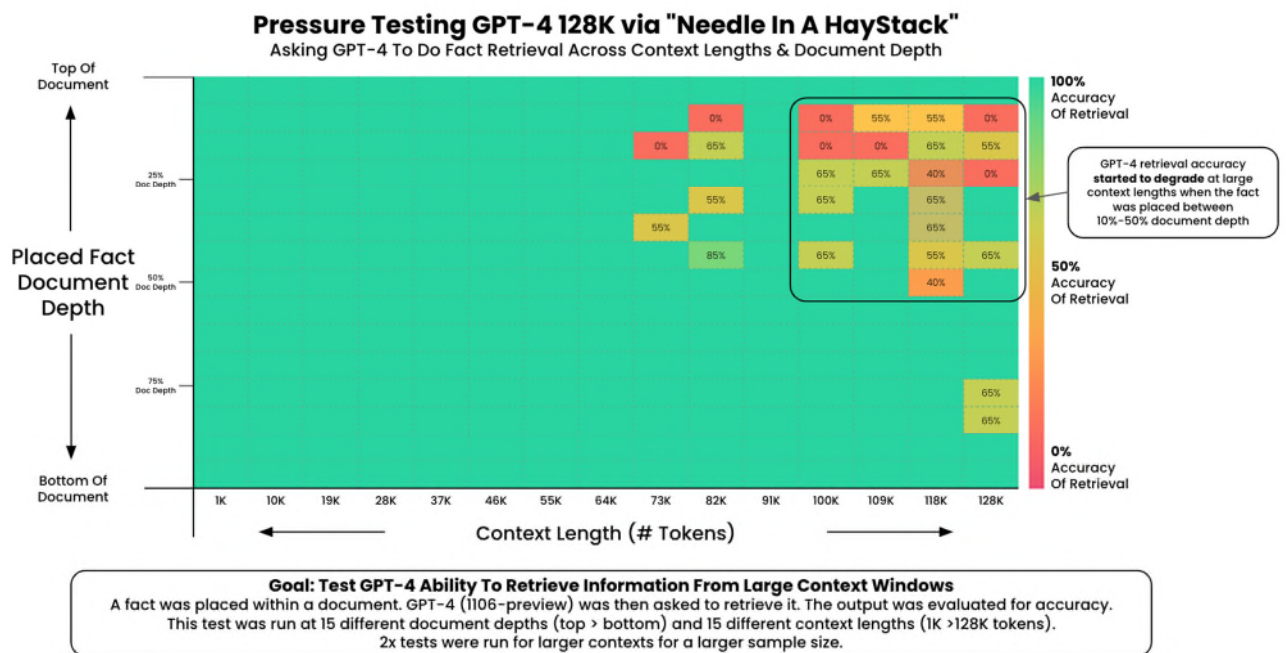


Figure 3. Pressure testing of Claude 2.1's context window via the "Needle In A Haystack" benchmark (Kamradt, 2023)

3 Prompt engineering

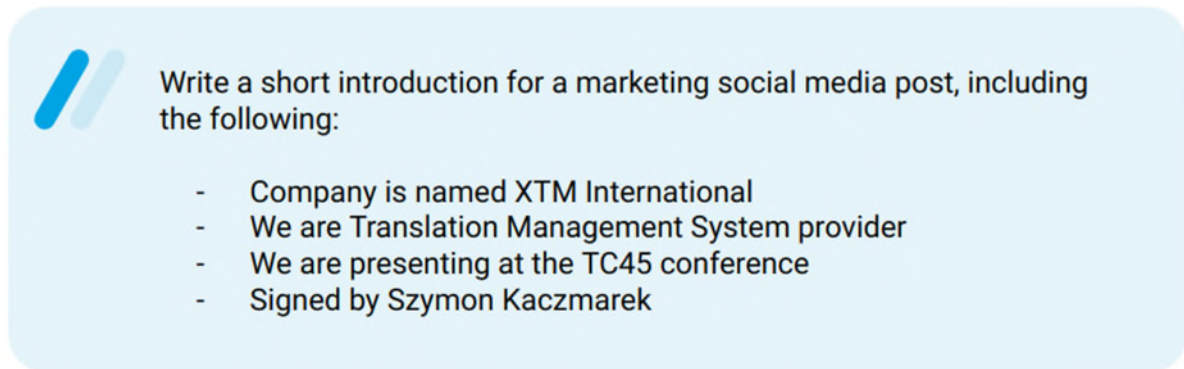
In the realm of machine learning, especially in natural language processing, prompt engineering has become a crucial field for leveraging the capabilities of LLMs. A prompt, in this concept, is an input sequence provided to an LLM to initiate and steer its generation process. The design of these prompts is a critical determinant of the quality and relevance of the model's output (Shin et al., 2020; Jiang, L., et al., 2021). A well-crafted prompt can turn a model from a collection of language patterns into a proactive entity capable of executing tasks that vary from completing texts to solving problems. Therefore, prompt engineering is more than just a technical task; it's a conversation where the user expresses their intent, and the model understands and replies accordingly.

The versatility of LLMs is, to a significant extent, a function of the ingenuity embedded in their prompts. A multitude of techniques have been developed to optimize this interaction, including basic instructions, in-context learning, role prompting, and more sophisticated methods like Chain of Thought (CoT) (Wei et al., 2022). Each of these techniques represents a unique approach to eliciting the best performance from an LLM.

3.1 Basic instructions

Basic instructions are the fundamental elements of prompt engineering, just like the primary colors, from which an infinite palette of interactions with LLMs can be achieved. The construction of these instructions requires precision and a nuanced understanding of the

model's capabilities and limitations. As shown in Figure 4, exemplary basic instruction for a marketing social media post might include essential details such as the company's name, the nature of the business, and key events.



Write a short introduction for a marketing social media post, including the following:

- Company is named XTM International
- We are Translation Management System provider
- We are presenting at the TC45 conference
- Signed by Szymon Kaczmarek

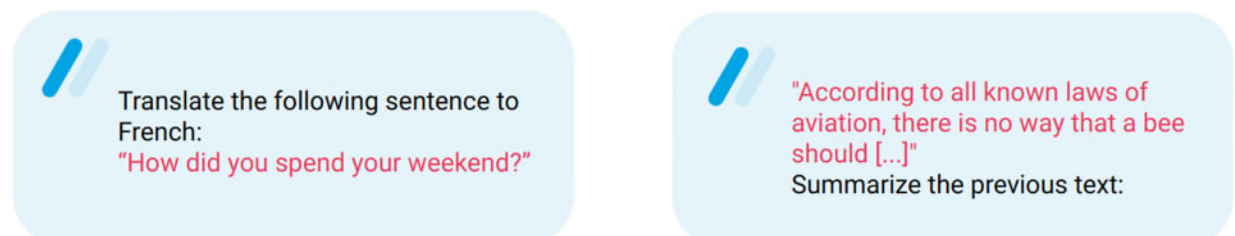
Figure 4. Exemplary basic instructions for a marketing social media post

This instruction is detailed enough to guide the LLM toward generating content that aligns with the company's marketing objectives, yet open-ended enough to allow for creative execution. Instructions should be crafted to be unambiguous, avoiding any misunderstandings that could lead to incorrect or irrelevant outputs.

These basic instructions set the stage for further, more complex forms of prompt engineering, thus mastering the formulation of them is crucial for leveraging the full potential of LLMs.

3.2 Primary content

The primary content of a prompt is the essence of the user's request to an LLM. It conveys the main task to be completed, whether it is translation, summarization, question answering, or any other application. For instance, a prompt requesting translation must specify the source sentence and the target language, as shown in Figure 5.



Translate the following sentence to French:
"How did you spend your weekend?"

"According to all known laws of aviation, there is no way that a bee should [...]"
Summarize the previous text:

Figure 5. Exemplary prompts with primary content and basic instructions

The effectiveness of LLMs in interpreting and executing these tasks has been the subject of extensive research, particularly in the field of translation. Studies comparing the translation

capabilities of LLMs have highlighted their potential to produce results that are increasingly comparable to human translators, although challenges remain in terms of consistency and handling of nuanced language (Hendy et al., 2023; Son and Kim, 2023; Jiao et al., 2023; Raunak et al., 2023). Methods like Adaptive Machine Translation using fuzzy matches, introduced in (Moslem et al., 2023) significantly improve translation quality.


In addition to translations, the primary content is also important in tasks such as summarization. The model's ability to distil complex text into a concise summary without losing critical information has significant implications for information retrieval and knowledge management.

3.3 In-context learning

In-context learning is a feature of the newest large language models (LLMs) that allows them to learn from examples given directly in the prompts. This is different from fine-tuning which requires bigger sets of labelled data and longer training times to learn tasks (Radford et al., 2018; Brown et al., 2020).

One-shot Learning: In one-shot learning, an LLM is given a single example within the prompt to understand what it needs to do. The model uses this example as a reference to generate a response to a new but similar task, demonstrating an immediate understanding from minimal information. For example, if one requires the output of LLM to be in a specific format, the prompt given in Figure 6 can be used.





Extract named entities from the corresponding texts below:

Text 1: Apple was founded in 1976 by Steve Jobs.
Output: Apple,1976,Steve Jobs

Text2: Jim bought 300 shares of Acme Corp. in 2006.
Output: Jim,300,Acme Corp.,2006

Figure 6. Example of a one-shot learning

Few-shot Learning: Few-shot learning extends this concept by providing the LLM with several examples. This approach helps the model to better understand the pattern or rule underlying the task.

In-context learning gains an advantage from the pre-trained knowledge of LLMs, allowing them to apply their extensive base of learned information to specific tasks with minimal additional input. This capability has profound implications for the speed and efficiency of deploying LLMs in practical applications. The selection and even permutation of examples is crucial, as they must be representative and informative enough to guide the model (Zhao et al., 2021; Lu et al., 2022). The model’s accuracy can be unstable depending on the prompt, as it tends to bias towards recent tokens, i.e. the model repeats answers that appeared at the end of the prompt. A good practice is to place *content-free* input as the last example (Zhao et al., 2021). Another cause of the high output variance is *Majority Label Bias* — LLMs are biased towards answers that are frequent in the prompt, i.e. training set is unbalanced. In a study (Min et al., 2022), it was demonstrated that labels provided in training examples can be random without negatively affecting performance across a variety of classification and multiple-choice tasks. What does matter is the distribution of the input text and the overall format of the sequence.

The few-shot learning approach seems to face challenges in complex tasks that require a deep understanding of context, such as translation or summarization. Reynolds and McDonell (2021) argue that few-shot prompts may not actually teach LLMs anything new but instead trigger the model to use its pre-existing knowledge base. Few-shot examples can however sometimes lead to worse performance as they may confuse the models, making them treat the examples as part of a narrative to be continued, rather than instructions to be followed.

3.4 Chain of Thought (CoT)

Chain of Thought (CoT) prompting is a strategic innovation in the field of LLMs that revolutionizes how these models approach problem-solving tasks. By prompting the model to reveal its reasoning in a stepwise manner, as shown in Figure 7, CoT mimics human cognitive processes, allowing for more transparent and interpretable decision-making (Wei et al., 2022).



Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more. How many apples do they have?

A:

Figure 7. Example of Chain of Thought prompting

The essence of CoT lies in its ability to break down complex problems into a series of logical steps, making it particularly effective for tasks that require multi-step reasoning, such as mathematics, causal reasoning, or common-sense justification. Kojima et al. (2023) have advanced the technique into the zero-shot domain, demonstrating that LLMs like GPT-3 can exhibit remarkable reasoning capabilities even without explicit examples, by simply being prompted to "think step by step" (Kojima et al., 2023). Similarly, researchers have found that when models are prompted to "think out loud," their performance on arithmetic word problems and other reasoning tasks improves markedly (Cobbe et al., 2021).

The technique has also been expanded into multilingual settings, where LLMs are prompted to articulate reasoning in languages other than English, thereby improving the cross-linguistic transfer of reasoning skills (Shi et al., 2022)

3.5 Role-Play Prompting

Role-play prompting stands out as a powerful tool in the field of LLMs, enabling these models to adopt personas and thus contextualize responses in a manner that mirrors the assumed roles. Kong et al. (2023) illustrate the efficacy of this approach, particularly under zero-shot conditions, where LLMs like ChatGPT and Llama 2 have shown significant improvements across various reasoning benchmarks when prompted to role-play.



From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students.

Xavier was 4 feet tall and grew 3 inches. Cole was 50 inches tall and grew 2 inches over the summer.

Q: What is the difference between Cole and Xavier's height now?

Figure 8. Example of Role-play prompting (Kong et al., 2023)

4 Conclusions

The emergence of Large Language Models (LLMs) marks a transformative phase in natural language processing, showcasing notable advances in model architecture and capabilities. This paper has explored the technical progress enhancing LLMs, the art of prompt engineering, and the challenges these technologies face.

We have witnessed a shift towards building larger models through significant parameter scaling, and the realization that efficiency and the use of strategic data are vital for achieving high performance. The ability to extend context windows has improved models' understanding, although this does not always translate to better performance in complex understanding tasks.

Prompt engineering has become crucial, evolving from simple instructions to sophisticated techniques like role-play and Chain of Thought, broadening LLMs' applicability and improving their accuracy and reliability.

All these advancements raise high expectations in the localization industry which is typically dealing with large volumes of multilingual texts. It is clear that companies are testing the new technology and trying to employ LLMs in real-world scenarios. Techniques involving machine translation, translation quality evaluation or automatic post-editing are more and more often deployed in production scenarios. The key to obtaining the highest quality and maximum gains from this technology is to understand LLMs and prompt them effectively.

References

- Abid et al. (2021). Persistent Anti-Muslim Bias in Large Language Models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.
- Agrawal et al. (2023). Do Language Models Know When They're Hallucinating References?, arXiv:2305.18248v2 [cs.CL]
- AI Index Steering Committee. (2023). AI Index 2023 Report. Stanford University, Human-Centered AI Institute. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf
- Almazrouei et al. (2023). The Falcon Series of Open Language Models, arXiv:2311.16867 [cs.CL]
- Anil et al. (2023). PaLM 2 Technical Report, arXiv:2305.10403 [cs.CL]
- Anthropic, (2023), Claude 2.1 Model Card Appendix, <https://www-cdn.anthropic.com/files/4zrzovbb/website/75639748080275c93d2ef9fc4239bdd111d7c234.pdf>
- Brown et al., (2020). Language Models are Few-Shot Learners, arXiv:2005.14165v4 [cs.CL]
- Cheng et al. (2023). Sources of Hallucination by Large Language Models on Inference Tasks, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2758–2774
- Chowdhery et al. (2022). PaLM: Scaling Language Modeling with Pathways, arXiv:2204.02311 [cs.CL]
- Ferrara (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. ArXiv, abs/2304.03738.
- Gallegos et al. (2023). Bias and Fairness in Large Language Models: A Survey. ArXiv, abs/2309.00770.
- Cobbe et al., (2021). Training Verifiers to Solve Math Word Problems, arXiv:2110.14168v2 [cs.LG]
- Guerreiro et al. (2023). Hallucinations in Large Multilingual Translation Models, Transactions of the Association for Computational Linguistics (2023) 11: 1500–1517.
- Hendy et al., (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv:2302.09210v1 [cs.CL]
- Hoffmann et al. (2022). Training Compute-Optimal Large Language Models. ArXiv, abs/2203.15556.
- Huang et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, arXiv:2311.05232v1 [cs.CL]
- Jiang, L., et al. (2021). How can we know what language models know? arXiv:1911.12543 [cs.CL]
- Jiao et al., (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine, arXiv:2301.08745v4 [cs.CL]

- Kamradt, (2023). LLMTest_NeedleInAHaystack. GitHub. Retrieved from https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main
- Kojima et al., (2023). Large Language Models are Zero-Shot Reasoners, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).
- Kong et al., (2023). Better Zero-Shot Reasoning with Role-Play Prompting, arXiv:2308.07702v1 [cs.CL]
- Kshetri (2023). Cybercrime and Privacy Threats of Large Language Models. IT Professional, 25, 9-13.
- Li et al. (2023). Privacy in Large Language Models: Attacks, Defenses and Future Directions. ArXiv, abs/2310.10383.
- Liu et al. (2023). Lost in the Middle: How Language Models Use Long Contexts, arXiv:2307.03172v3 [cs.CL]
- Lu et al., (2022). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity, arXiv:2104.08786v2 [cs.CL]
- Luccioni et al., (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. Journal of Machine Learning Research 24 (2023) 1-15
- Min et al., (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, arXiv:2202.12837v2 [cs.CL]
- Moslem et al., (2023). Adaptive Machine Translation with Large Language Models, arXiv:2301.13294v3 [cs.CL]
- Navigli et al. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. ACM Journal of Data and Information Quality, 15: 1 - 21.
- OpenAI et al. (2023). GPT-4 Technical Report, arXiv:2303.08774 [cs.CL]
- Pan et al. (2023). On the Risk of Misinformation Pollution with Large Language Models. , 1389-1403.
- Radford et al., (2018) Improving language understanding by generative pretraining. Technical Report.
- Raunak et al., (2023). Leveraging GPT-4 for Automatic Translation Post-Editing, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12009–12024 December 6-10, 2023
- Reynolds and McDonell, (2021), Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm, arXiv:2102.07350 [cs.CL]
- Shi et al., (2022). Language Models Are Multilingual Chain-of-thought Reasoners, arXiv:2210.03057v1 [cs.CL]
- Shin, R., et al. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4222–4235
- Son and Kim, (2023). Translation Performance from the User’s Perspective of Large Language Models and Neural Machine Translation Systems. Information 2023, 14, 574.

- Tonmoy et al. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, arXiv:2401.01313v3 [cs.CL]
- Touvron et al. (2023a). LLaMA: Open and Efficient Foundation Language Models, arXiv:2302.13971 [cs.CL]
- Touvron et al. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models, arXiv:2307.09288 [cs.CL]
- Vaswani et al. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA
- Wei, J., et al. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. 36th Conference on Neural Information Processing Systems (NeurIPS 2022).
- Zhang et al. (2023). Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models, arXiv:2309.01219v2 [cs.CL]
- Zhang et al. (2023). How Language Model Hallucinations Can Snowball. arXiv:2305.13534v1 [cs.CL]
- Zhao et al., (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models, arXiv:2102.09690v2 [cs.CL]
- Zhong, L., & Wang, Z. (2023). A Study on Robustness and Reliability of Large Language Model Code Generation. ArXiv, abs/2308.10335.

Envisioning the Post-Editor's Workstation: A Backward Glance and a Glimpse into the Future

Marie Escribe

Universitat Politècnica de València
& LanguageWire, Spain
mcescrib@doctor.upv.es

Miguel Ángel Candel-Mora

Universitat Politècnica de València,
Spain
mcandel@upv.es

Abstract

The improvement in quality and precision of Machine Translation (MT) outputs has captured the attention of both the academic research community and industry professionals. This has led to a focus on optimising Post-Editing (PE) tasks to enhance the training of translation engines and improve the workflow of post-editors. However, translation environments, PE tasks, and quality assessment processes are still dependent on Computer-Assisted Translation (CAT) tools. The neural MT paradigm requires a workstation specifically designed for PE tasks, as the requirements for PE differ from those of translation, particularly in areas such as PE process analysis, PE guidelines, and PE skill sets. The aim of this paper is to establish a path for the development of a workstation specifically customised for the needs of PE tasks. To achieve this objective, an analysis of the main Translator's Workstation initiatives has been conducted, via literature review, to identify the functions that were initially developed to enhance the translator's workflow and, subsequently, employed for the development of present-day CAT tools. Then, we identified the common requirements in PE tasks, which served as the basis for presenting a prototype of a tool tailored to meet the needs of PE.

1 Introduction

Given the quality of Machine Translation (MT) outputs, particularly since the advent of deep learning, this technology has become an integral part of the translation workflow. Today, MT is commonly integrated into Computer-Assisted Translation (CAT) tools, and language professionals typically post-edit MT outputs in these environments. Nevertheless, CAT tools were not designed to support Post-Editing (PE) as a main task. The early proposals of a Translator's Workstation were made at a time when today's MT performance was difficult to foresee. Moreover, the integration of MT into CAT did not seem to be accompanied by specific adaptations of the CAT environment to this task. Based on these observations, one could raise doubts regarding the suitability of CAT tools for PE.

The objective of this paper is precisely to examine this question. Then, the first proposals of Translators' Workstations are reviewed, and PE requirements are examined. Several innovative solutions are also presented. Based on this analysis, the present research attempts to envisage a PE Workstation, which is presented in a prototype.

2 Theoretical Background

2.1 Conceptualisation of the Translator's Workstation

The evolution of translation technology has regularly oscillated between the desire to achieve fully automated translations and the desire to provide translators with the best possible help to do their job. Warren Weaver's 1947 proposal, according to which computers could automate translation based on the code breaking principles of World War Two, is known as the main origin of MT (Hutchins, 2005). A wave of optimism followed this initial proposal, culminating in a demonstration of the Georgetown-IBM system in 1954, known as the first public demonstration of MT (Hutchins, 2014). However, this enthusiasm was short-lived, since in 1966 the Automatic Language Processing Advisory Committee (ALPAC) report concluded that MT research was not destined for a promising future, and advised turning instead to translation aids, such as those used by the Federal Armed Forces Translation Agency or the Terminological Bureau of the European Coal and Steel Community, which made it possible to achieve significant productivity gains.

Even before the ALPAC report was published, the idea of using technology as a support instead of aiming for entire automation had already been put forward. Indeed, Bar-Hillel questioned the objective of reaching fully automatic translations and called instead for a "machine-post-editor partnership" (Bar-Hillel, 1960, p.97). Similarly, Licklider described a "symbiotic partnership" between humans and machines (Licklider, 1960, p.4).

Following the recommendations of the ALPAC, translation aids (known as translator's workstations or workbenches) began to attract more attention in the 1970s and 1980s, when these tools (or some of their components) were conceptualised (Hutchins, 1998). These proposals mostly revolved around a central focus: gathering all tools and resources that translators need to complete their work in a single environment.

Among relevant initiatives, Krollmann (1971) described linguistic data banks, Lipmann (1971) imagined a workstation supporting several text processing actions and access to remote terminological databases, and Arthern (1978) suggested a text retrieval mechanism to reuse similar translations and avoid repetitive work.

In the 1980s, Kay and Melby also presented detailed proposals for a translator's workstation, with both practical descriptions and strong theoretical implications. Kay described what he named the Translator's Amanuensis, a system integrating various functionalities designed to assist translators, such as suggestions retrieved from previous translations, dictionary look-up and morphology-aware processing (Kay, 1980). Beyond its technical dimension, the Translator's Amanuensis is also a theoretical concept in which man-machine collaboration plays a central role. According to Kay, although this system could gradually take over certain tasks, it should remain supervised by the human translator, as the main objective of this tool would be "to help increase his productivity and not to supplant him" (Kay, 1980, p.18). Similarly, Melby (1982) suggested a workstation with three different levels of assistance, ranging from terminological support only, to integration of MT, and envisaged that translators could switch between these levels as needed, since the different segments inside a text would not necessarily require the same degree of assistance.

These early conceptualisations paved the way for Computer-Assisted Translation (CAT) tools as we know them today, with text processing functionalities, access to Term Bases (TBs)

and making use of past translations with Translation Memories (TMs). The first commercial workstations were launched in the 1980s and 1990s (Hutchins, 1998) and today, translators can choose from a plethora of CAT tools.

Nonetheless, despite a significant slowdown following the ALPAC report, MT research continued to grow. Ten years after the publication of the report, the Pan American Health Association started to develop SPANAM, a Spanish-English MT model which became an integral part of the institution's translation activities (Vasconcellos and León, 1985). Similarly, the METEO system, specifically designed by the University of Montreal to translate weather forecasts from English into French, was operational since 1977 (Thouin, 1982). Following this progressive resurgence, MT systems started to yield more and more encouraging results, as the field moved from one paradigm to another.

Today, Neural MT (NMT) represents state-of-the-art technology and is known to generate outputs of high quality. With the increasing amount of content to be translated in today's globalised world, and given the efficiency of deep learning methods, integrating NMT in translation workflows became a standard practice. Therefore, a typical setup today consists of post-editing MT output within a CAT environment.

2.2 Post-Editing Requirements

CAT tools were originally designed to optimise translation, not necessarily PE. Therefore, it is arguable whether current CAT tools are well suited to PE.

A comparative analysis of the translation and PE processes offers a starting point to answer this question. In the field of Translation Process Research, there seems to be a consensus that translation can be divided into three phases, although the names and boundaries between each phase may differ from one model to another. Sager (1994) proposed to divide the translation process into reading comprehension, translation and revision. Similarly, do Carmo (2017) distinguished the following phases: orientation (planning and reading), drafting (generating the target text) and self-revision (during which the translation is checked). Instead of a linear activity, translation is rather an iterative process, as translators are likely to edit repeatedly their draft until they are satisfied with their rendition (*ibid.*).

do Carmo (2017) provided a thorough analysis of the PE process, and defined it based on the three phases of translation: orientation (consulting the source and target texts), drafting (checking the MT output and amending it when necessary) and self-revision (checking the translation). Although this establishes a common reference, the use of "drafting" for the PE process might lead to confusion, as drafting is typically understood as writing a first version of a text, which does not appear to be an accurate description of PE, as the target text is already present. PE is typically described as a process of correcting errors found in MT outputs (as further discussed in the following paragraphs), and it would therefore be tempting to use the term "correction" instead of "drafting". However, "correction" would imply that the raw MT systematically contains errors. This assumption is not always accurate, especially given the high quality of NMT, which sometimes leads to directly accepting some MT suggestions without performing any amendment. Based on this reflection, this phase could be named "validation", but this term does not constitute an exact representation of the reality of PE either, as it relies on the hypothesis that MT outputs can be trusted. Instead, a more balanced option seems more reasonable, hence we suggest calling this phase "adjustment". This term is more neutral and describes the task of applying changes to obtain a result which fits the requirements, which is a more accurate definition of this central phase in PE.

The study of PE competence models and guidelines also sheds light on PE needs. In comparison to translation, for which competence models have long been explored, PE has emerged more recently. However, PE models have been introduced in the last few years. While it is generally acknowledged that PE requires the same basic competences as translation, certain skills are specific to PE. Among relevant competence models, Nitzke and Hansen-Schirra (2021) distinguished three pillars: error handling, MT engineering and consulting. These constitute three different areas of specialisations that post-editors may pursue. The dimension which is the most related to “practical” PE is error handling, which is in turn subdivided into error spotting, classification and correction. Similarly, Ginovart Cid (2021) identified three core skills for practical PE: error spotting, decision making and application of guidelines.

As far as guidelines are concerned, although these may differ significantly depending on the scenario, certain patterns can be found. Hu and Cadwell (2016) compared five sets of PE guidelines and identified overlaps and discrepancies. In the case of full PE, most guidelines were found to share the same vision regarding accuracy, terminology and grammar, but express different views on style. Most guidelines also encourage post-editors to use as much of the MT output as possible, thus avoiding preferential changes (Hu and Cadwell, 2016; Nitzke and Hansen-Schirra, 2021). Moreover, the TAUS guidelines (TAUS, 2016) and ISO 1858 standard for PE services (ISO, 2017) are commonly recognised as references in the industry. They both state that the output of a PE task should be a translation which is accurate, linguistically correct, and compliant with the relevant requirements.

The main difference between translation and PE therefore lies primarily in the fact that in one case, the translator starts composing the target text from scratch, whereas in the other case, a translation suggestion is already available, and the post-editor can either accept it, amend it, or reject it (and then proceed to retranslate in the last case). Consequently, while thinking of various translation options and typing are the main tasks in translation, PE revolves around detecting errors and correcting them via the four editing actions, namely deleting, inserting, moving and replacing (do Carmo, 2017).

This observation should be refined, as the central focus on editing an existing suggestion, as opposed to generating a draft, also occurs when translating with TMs. In the case of fuzzy matches, the focus is also on adapting suggestions, which often takes the form of editing actions. In this sense, Mossop (2020) pointed out that the work of a translator using TMs is closer to that of a revisor/editor than that of a composer. Arguably, this phenomenon is exacerbated in the case of PE.

Nevertheless, editing TM fuzzy matches and MT outputs are radically different tasks. Fuzzy matches come with a matching score, which indicates the difference between the entry stored in the TM and the current segment, thus also providing an estimation of the amount of editing required. Moreover, the differences between the current segment and the retrieved TM match are typically highlighted in CAT tools for an easier visualisation. In the case of MT, despite its central role, error identification is not supported by any specific CAT tool component (apart from certain quality checks).

3 Current Technological Landscape for Post-Editing Environments

While it remains true that CAT tools did not go through drastic changes since the integration of MT, several proposals have been put forward. This section aims at exploring suggestions made to optimise the PE environment and compare them with the real working conditions of post-editors using commercial systems.

3.1 Proposals for Optimising Post-Editing

Several user studies have investigated user needs in terms of PE and a few attempts have been proposed to adapt these tools to PE. Moorkens and O'Brien (2017) studied user attitudes towards PE tools and found that confidence scores and dynamic adaptation of MT outputs were deemed important to participants, among other features.

Furthermore, several proposals of PE-centric technologies have been introduced. For example, certain systems rely on interactive MT to provide prefix-constrained suggestions, which allow for dynamic adaptations based on the translator's input (Wuebker et al., 2016; Santy et al., 2019; Peris and Casacuberta, 2019). The display of Automatic Post-Editing (APE) suggestions in PE interfaces has also been suggested, as done in the OpenTIPE tool (Landwehr, Steinmann and Mascarell, 2023).

IntelliCAT (Lee et al., 2021) offers Quality Estimation (QE) at the sentence level (in the form of confidence scores) and at the word level (highlighting potential issues, such as incorrect words and locations of missing words). This tool can also provide alternative translations upon the translator's request.

Other systems revolve around more diverse interaction modalities. This is the case of the tool introduced by Teixeira et al. (2019), which supports touch and speech input, and the Multi-Modal Post-Editing (MMPE) interface, which includes voice commands as well as handwriting and touch interactions (Herbig et al., 2020).

Herbig et al. (2019) suggest relevant directions to leverage synergies between humans and technology with a view to optimising the PE environment. After identifying error detection and correction as the core task in PE, they proposed various solutions to further support this. Their suggestions include three types of assistance: QE, source-MT alignments for a fast comparison, and colour coding to detect similarity between source segments and their matches retrieved from a TM.

Alonso and Nunes Vieira (2017) described the "Translator's Amanuensis 2020" (TA2020) based on the PE requirements identified in the literature. The TA2020, in the expert level (i.e., post-editor facing) would offer a set of modern features. Namely, a knowledge feature would display relevant information (i.e., identifying keywords and providing definitions or images from related sources) to post-editors upon request or when the tool detects that they might require support; an effort prediction feature would provide an estimation of the remaining amount of work to be completed; and a feature 3D visualisation would overcome decontextualization issues by displaying the source and target content in two layers, with the target text in the foreground.

Several proposals (Alonso and Nunes Vieira, 2017; Herbig et al., 2019; O'Brien, 2021) suggested monitoring the cognitive load to detect moments in which the post-editor encounters difficulties and provide tailored assistance (based on gaze data from eye tracking). This approach is also reminiscent of the principles behind *Escriba* (Porto Veloso, 2013), a CAT tool with an adaptive user interface, which was developed with the intention of predicting user actions based on behavioural patterns.

The integration of further functionalities into sometimes already crowded CAT interfaces is however likely to be a challenging task. In particular, the abundance of translation suggestions coming from different sources (TM, MT, APE) might be overwhelming, and it is therefore crucial to find an optimal way of presenting this information to the post-editor (Herbig et al., 2019). This may clash with the common preference to have a lean UI (Moorkens and O'Brien,

2017), and therefore involves finding an optimal balance between feature richness and simplicity.

Another type of balance should also be considered: that between maintaining interfaces similar to what translators are currently using to avoid abrupt changes, and radically modifying tools and relying on users to adapt to new functionalities. As regards this aspect, Alonso and Nunes Vieira (2017) contended that post-editors would have to adapt to different ways of visualising information. Arguably, the transition to PE-adapted interfaces could also be envisaged gradually, introducing a few features at a time to avoid hasty disruptions. Moreover, another possible solution to this issue lies in personalisation: customisable functionalities appear as a recurring need (Moorkens and O'Brien, 2017; Alonso and Nunes Vieira, 2017; O'Brien, 2021) and would allow users to configure the UI based on their preferences.

3.2 The Reality of the Market

It appears that despite the integration of MT and the fact that PE is mostly performed in a CAT environment, CAT tools are not yet adapted to this shift (Herbig et al., 2019). While certain components (such as TBs) remain valuable when MT is used (*ibid.*), translation and PE seem bound to the same limitations of CAT – one of the most frequently mentioned being the decontextualization phenomenon resulting from segmentation (Candel-Mora, 2015). Beyond this observation, performing a task in an environment which was designed for a different – albeit similar – activity suggests that there is room for optimisation.

Moreover, PE-specific features seem to be more frequently discussed in research circles than implemented in commercial applications. Speech input features can now be used in several CAT tools (e.g., Hey memoQ dictation add-on, Web Speech functionality in Wordfast Anywhere), and the same applies for multimodal functionalities such as the Text To Speech plug-in for Trados. However, it can be argued that external dictation integrations are not as powerful as embedded functionalities, since the latter could support more powerful interactions, notably via voice commands as in the case of MMPE.

More recently, Large Language Models (LLMs) began to be incorporated into CAT tools. For example, GPT-4 was integrated into Matecat¹ in May 2023 to allow for in-context search. Although some tools already offered to open pre-selected web pages upon selecting words and pressing a button (e.g., memoQ Web Search), Matecat's AI assistant provides results based on the source content context. Since August 2023, AI Actions (namely, rephrasing, shortening, translating with GPT-3.5 and fixing punctuation and grammar) also became available in SmartCat².

Above all, it appears that today's technology has drifted away from the original ideas of the Translator's Workstation formulated in the 1980s. Admittedly, a significant part of the first proposals has been implemented, since TMs and TBs remain the cornerstone of modern CAT tools. This should however be nuanced, as some advanced functionalities were conceptualised in the first proposals but have remained absent or have been integrated only recently in CAT tools. Morphology-aware processing (mentioned by both Kay [1980] and Melby [1982]) is a

¹ Source: <https://translated.com/matecat-gpt-4> (Accessed: 25/9/2023).

² Source: <https://www.smartcat.com/release-notes/> (Accessed: 25/9/2023).

good example. In 2022, Matecat announced the release of a new glossary³, which uses matching techniques capable of recognising declensions⁴. Similarly, the terminology recognition mechanism in Smart Editor (LanguageWire's CAT tool) will adopt a lemmatisation approach before 2024. Although this is a significant enhancement, this type of morphology-aware addition could be extended to other features beyond terminology recognition, in particular concordance and search features.

Similarly, confidence scores, which Melby (1982) had described as a “self-evaluation metric” for MT, are not commonly integrated in most commercial systems. One notable exception is Phrase, where confidence scores are displayed for each MT segment⁵.

One could argue that only the practical suggestions sustaining the concept of the Translator's Workstation have been implemented, but the theoretical proposal behind it has been neglected – if not forgotten – over the years. The human-machine partnership described by Bar-Hillel and Licklider in 1960 seems far from today's reality, whereby the use of MT and CAT tools are rather imposed on translators. The human-centric proposals of Kay and Melby in the 1980s, which aimed to empower translators by augmenting their capabilities and giving them full control over which level of assistance they required, now seems to be a dream of the past.

Despite some attempts to elicit PE needs and to adapt CAT tools, it appears that there is still no consensus regarding what the ideal PE environment should be. While research seems to focus on improving MT and APE systems, comparatively limited attention is paid to PE tools. However, do Carmo (2017, p.199) pointed out that “the current challenge for PE is not so much on improving the quality of the MT output, but on giving translators proper conditions for their job, including PE”.

4 Proposal

Based on the observations above, it appears relevant and necessary to rethink the optimal PE environment. As discussed, several recommendations have been put forward, however, to the best of our knowledge, they have not been gathered into a unique proposal.

For example, some tools now come with AI assistants, but do not provide translation alternatives or confidence scores. Some do offer alternatives and confidence scores, but do not support multimodal interactions (Lee et al., 2021). Other tools are centred on multimodality and also support QE (Shenoy et al., 2021), but not alternative translations. Most of these features are not available in the well-known commercial platforms. All systems also seem to rely on sentence segmentation. No tool allows the user to specify the desired level of assistance, nor detects moments in which the translator might need extra support.

In the light of the above discussion, and with the aim of designing a PE environment adapted to the needs of post-editors, this study attempts to envision a translation tool supporting these PE-specific features in the same interface. The resulting proposal is presented below.

³ Source: <https://guides.matecat.com/matecat-release-notes> (Accessed: 25/9/2023).

⁴ Source: <https://guides.matecat.com/work-with-the-glossary-while-translating> (Accessed: 25/9/2023).

⁵ Source: <https://phrase.com/blog/posts/mt-quality-estimation/> (Accessed: 25/9/2023).

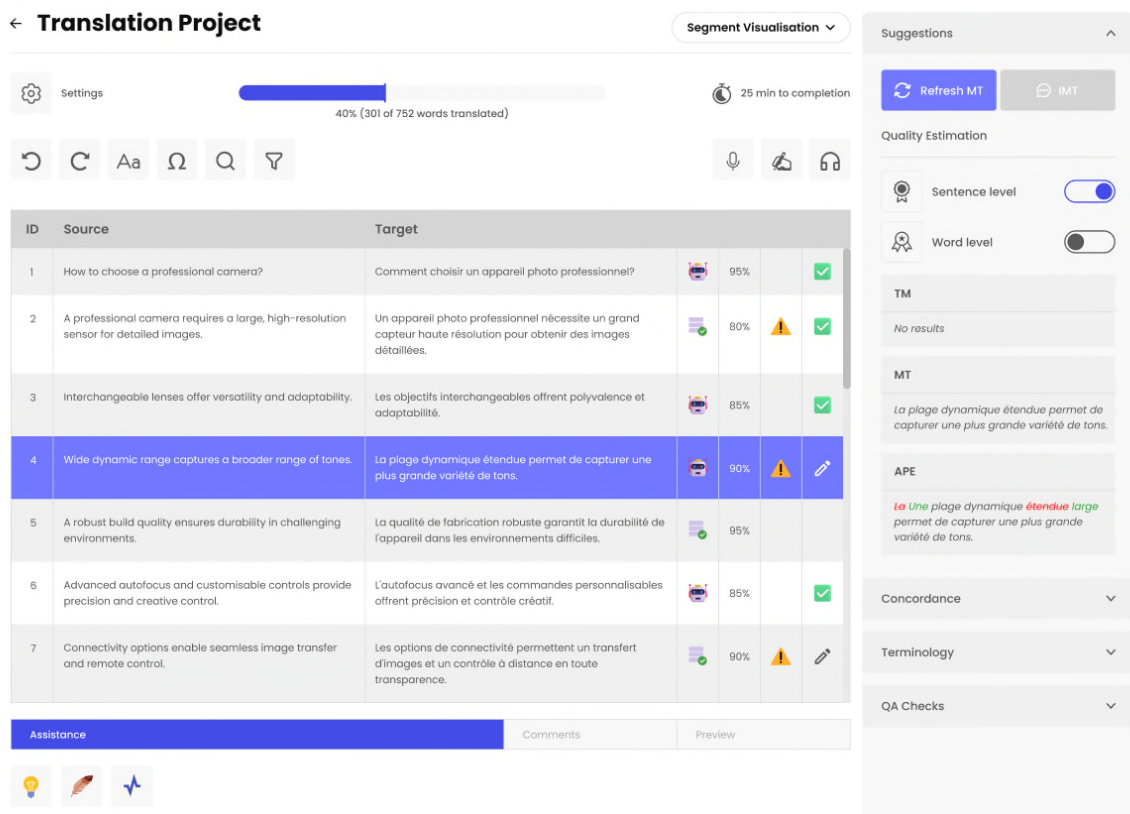


Figure 1. Proposal of a Post-Editor's Workstation

The features of this PE environment are described below. They are divided into five main categories: translation suggestions (displaying and interacting with the different translation options provided), assistance (support for decision making), interaction modalities (forms of input and output supported in the tool), visualisation (in-context view of the texts) and performance analysis (examination of productivity).

- **Translation Suggestions**

- TM: integration with regular TMs.
- MT: MT outputs, with the possibility of refreshing the output (i.e. regenerating the text considering the edits already implemented).
- IMT: the post-editor can choose whether to work with static or interactive MT suggestions.
- APE: APE suggestions, in which the differences between the original MT output and the correct version are highlighted.
- In-segment alignments: clicking on a word in one side of a segment highlights the corresponding word in the other side.

- In-segment alternative translations: right-clicking on a word in the target side displays a list of alternative translation suggestions for this word, based on the context.
- ***Assistance***
 - QE: the post-editor can activate and deactivate QE information at two different levels.
 - Sentence level: sentence-level confidence scores are displayed at the segment level.
 - Word level: word-level quality indicators are displayed in the form of a colour code for each word based on its estimated quality.
 - Knowledge feature: the post-editor can send selected text fragments to this feature, which will output definitions or explanations, possibly including images.
 - Rewriting: the post-editor can benefit from rewriting suggestions for pre-defined situations through LLM integrations (for example, rewriting in a different style or shortening a translation).
- ***Interaction modalities***
 - Speech input: the post-editor can dictate their edits (e.g. selecting a fragment of text and dictating the text which should be used to replace it instead of typing) and use voice commands to navigate in the tool.
 - Text to speech: the post-editor can use this feature to listen to the source or target text.
 - Touch input: the post-editor can use this functionality to enter text via handwriting or to perform touch interactions, such as dragging and dropping text fragments.
- ***Visualisation***
 - Document view: the post-editor can switch from the segment visualisation to in-document visualisation, in which the source and target documents are displayed side by side in their original format. The target text can be modified directly in this view, and the segments being modified are also highlighted in the source for reference.
- ***Performance analysis***
 - Effort prediction: based on performance analysis in previous assignments, an expected time to completion is indicated. After a job is finished, this feature can also serve for further analysis by comparing the expected and actual productivity.
 - Attention monitoring: an analysis of behavioural patterns (possibly including eye tracking data) can help detecting drops in productivity (and display a warning) or moments in which the post-editor may require a certain type of assistance (and propose this assistance automatically).

- Post-task feedback: once the PE task is completed, the tool can provide immediate feedback based on the proportion of accepted, amended and rejected segments. While accepted segments can easily be spotted, the difference between amending and rejecting can be more delicate to find based on the edit distance, however tracking user actions (especially deleting text) can help disambiguate the boundaries between amending and rejecting.

It should be noted that this proposal is first and foremost a work of ideation, aimed at assembling various recommendations within a single tool. The prototype provided herein is presented as an example of what future PE tools could potentially be like and aims to serve as a starting point for further discussion.

5 Conclusions

Compared with translation, PE has various specificities, particularly in terms of process, competence and guidelines. Therefore, adapting current translation environments by including PE-specific features seems a relevant approach to provide better assistance to post-editors. The Translator's Workstation revolved around the concepts of human-machine partnership and augmented translation. Such concepts should continue to guide the development of translation tools today, especially given the shift to PE. While several suggestions have already been introduced, these tend to come from various sources and are still not implemented in most commercial systems.

Based on these observations, a prototype of a PE environment was introduced. It offers various PE-specific features designed to enhance the experience of post-editors by providing assistance, adapted translation suggestions, more diverse interaction modalities and visualisation modes, and a performance analysis module. At this stage, this work constitutes mostly a theoretical contribution and seeks to rethink PE environments.

6 Limitations and Future Work

As a next step in this study, it will be paramount to test the prototype introduced herein with professional post-editors. A validation round should indeed be conducted to gather feedback on the PE-specific features included in the prototype. The primary objectives should be to find out whether post-editors would imagine using these functionalities in their daily work, identify any suggestions for improvements, and discuss other potential additions for an optimised PE environment. Based on the feedback received, if deemed necessary, a second version of the prototype could be designed, and tested again in a second iteration.

In addition, user studies would be highly beneficial to measure the impact of PE-specific features on productivity, quality and satisfaction in working conditions similar to real-life PE assignments. To date, several functionalities remain out of focus. While Moorkens and O'Brien (2017) found that users would be interested in seeing MT confidence scores, a study conducted two years previously revealed that displaying PE effort indicators did not influence the PE effort (Moorkens et al., 2015). This finding calls into question the relevance of confidence score as a support for post-editors. Nonetheless, it could be argued that this study was conducted at the beginning of NMT, and that the quality of MT outputs has improved since 2015, which could also affect the perceived usefulness of confidence scores.

Finally, as pointed out by Moorkens and O'Brien (2017), it is crucial to recall that CAT tools already have several flaws and sometimes fail to meet users' requirements in translation scenarios (i.e. even when MT is not used). Therefore, any endeavour to adapt CAT tools to PE is a delicate exercise, as the foundation to be adapted should perhaps first be improved.

Acknowledgements

This research is conducted as part of an industrial Ph.D. agreement between the Universitat Politècnica de València and LanguageWire.

References

Alonso, Elisa, and Lucas Nunes Vieira. 2017. The Translator's Amanuensis 2020. *The Journal of Specialised Translation*, July, pages 345-61.

ALPAC. 1966. *Language and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council.

Arthern, Peter J. 1978. Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint. In *Proceedings of Translating and the Computer*.

Bar-Hillel, Yehoshua. 1960. The Present Status of Automatic Translation of Languages. *Advances in Computers*, 1, pages 91-163.

Candel-Mora, Miguel Ángel. 2015. Comparable Corpus Approach to Explore the Influence of Computer-Assisted Translation Systems on Textuality. *Procedia-Social and Behavioral Sciences*, 198, pages 67-73.

do Carmo, Félix Emanuel Martins. 2017. Post-Editing: A Theoretical and Practical Challenge for Translation Studies and Machine Learning. Ph.D. Thesis, Universidade do Porto.

Ginovart Cid, Clara. 2021. The Need for Practice in the Acquisition of the Post-Editing Skill-Set: Lessons Learned from the Industry. Ph.D. Thesis, Universitat Pompeu Fabra

Herbig, Nico, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019. Integrating Artificial and Human Intelligence for Efficient Translation. Preprint.

Herbig, Nico, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. MMPE: A Multi-Modal Interface for Post-Editing Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1691-1702.

Hu, Ke, and Patrick Cadwell. 2016. A Comparative Study of Post-Editing Guidelines. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 34206-353.

Hutchins, John. 1998. The Origins of the Translator's Workstation. *Machine Translation* 13 (4), pages 287-307.

Hutchins, John. 2005. The History of Machine Translation in a Nutshell.

ISO. 2017. ISO 1858 Standard. Translation Services – Post-editing of Machine Translation Output – Requirements.

Kay, Martin. 1980. The Proper Place of Men and Machines in Language Translation. *Xerox PARC CSL-80-11*.

Krollmann, Friedrich. 1971. Linguistic Data Banks and the Technical Translator. *Meta: Translators' Journal*, 16(1-2), pages 117-124.

Landwehr, Fabian, Thomas Steinmann, and Laura Mascarell. 2023. OpenTIPE: An Open-Source Translation Framework for Interactive Post-Editing Research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 208-216.

Lee, Dongjun, Junhyeong Ahn, Heesoo Park, and Jaemin Jo. 2021. IntelliCAT: Intelligent Machine Translation Post-Editing with Quality Estimation and Translation Suggestion. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 11-19.

Licklider, Joseph C. R. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, 1, pages 4-11.

Lippmann, Erhard O. 1971. An Approach to Computer-Aided Translation. *IEEE Transactions on Engineering Writing and Speech*, 14(1), pages 10-33.

Melby, Alan K. 1982. Multi-Level Translation Aids in a Distributed System. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, pages 215-220.

Moorkens, Joss, and Sharon O'Brien. 2017. Assessing User Interface Needs of Post-Editors of Machine Translation. In *Human Issues in Translation Technology*, pages 127-148. Routledge.

Moorkens, Joss, Sharon O'Brien, Igor A. L. da Silva, Norma B. de Lima Fonseca, and Fabio Alves. 2015. Correlations of Perceived Post-Editing Effort with Measurements of Actual Effort. *Machine Translation* 29 (3), pages 267-284.

Nitzke, Jean, and Silvia Hansen-Schirra. 2021. *A Short Guide to Post-Editing* (Volume 16). Language Science Press.

O'Brien, Sharon. 2021. Translation, Human-Computer Interaction and Cognition. *The Routledge Handbook of Translation and Cognition*, pages 376-388. Routledge.

Peris, Álvaro and Francisco Casacuberta. 2019. Online Learning for Effort Reduction in Interactive Neural Machine Translation. *Computer Speech & Language* 58(1), pages 98-126.

Porto Veloso, Pablo. 2013. *Escriba*, an Adaptive Web CAT Tool. MA Dissertation. University of Dublin, Trinity College.

Santy, Sebastin, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive Neural Machine Translation Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 103-108.

Shenoy, Raksha, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2021. Investigating the Helpfulness of Word-Level Quality Estimation for Post-Editing Machine Translation

Output. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10173-10185.

TAUS (Isabella Massardo, Jaap van der Meer, Sharon O'Brien, Fred Hollowood, Nora Aranberri, Katrin Drescher). 2016. *Machine Translation Post-Editing Guidelines*.

Teixeira, Carlos S. C., Joss Moorkens, Daniel Turner, Joris Vreeke, and Andy Way. 2019. Creating a Multimodal Translation Tool and Testing Machine Translation Integration Using Touch and Voice. *Informatics* 6 (1).

Thouin, Benoit. 1981. The Meteo System. In *Proceedings of Translating and the Computer: Practical Experience of Machine Translation*, pages 39-44.

Vasconcellos, Muriel, and Marjorie León. 1985. Spanam and Engspan: Machine Translation at the Pan American Health Organization. *Computational Linguistics* 11 (2-3), pages 122-136.

Wuebker, Joern, Spence Green, John DeNero, Saša Hasan, and Minh-Thang Luong. 2016. Models and Inference for Prefix-Constrained Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66-75.

Post-Editing Machine Translation Beyond the Binary: Insights into Gender Bias and Screen Activity

Manuel Lardelli

Department of Translation Studies,

University of Graz

manuel.lardelli01@gmail.com

Abstract

Machine Translation (MT) knowingly suffers from gender bias and research in this field often focuses on the tendency of MT systems to overuse masculine forms or reproduce gender stereotypes. Few studies address the challenges of non-binary representation. Non-binary, gender-fair approaches are numerous and language-specific, thus debiasing MT is particularly arduous. For this reason, I propose a case study on gender-fair post-editing. Six professional translators were asked to post-edit three English-to-German machine translations. For each text, they were instructed to utilize a different gender-fair language (GFL) approach, i.e., gender-neutral rewording, gender-inclusive characters, and neosystems. The focus of this study is on bias in the machine-translated outputs and on the GFL post-editing process which was reconstructed with screen recordings. Findings from the analysis of the machine translations show that DeepL systematically misgenders and erases non-binary identities. The analysis of the screen recordings suggests substantial differences among GFL approaches in screen activity, with a high variability among participants. These results provide insights into the limitations of specific GFL strategies and highlight how there is no one-fits-all solution to non-binary representation. Cross-fertilization among disciplines such as translation studies and computational linguistics as well as participatory approaches are needed to achieve gender-fairer machine translation.

Content warning: this paper contains examples of misgendering and erasure that could be offensive and triggering to trans and non-binary individuals.

1 Introduction

Machine Translation (MT) knowingly suffers from gender bias. Research in this field generally focuses on the reproduction of gender stereotypes and the tendency of MT systems to overuse masculine forms. Furthermore, numerous debiasing techniques have been proposed (Savoldi, Gaido, Bentivogli, Negri, & Turchi, 2021). Most of the studies, however, still neglect non-binary individuals (Lardelli & Gromann, 2023a) in spite of their increased visibility in the last few years. This is a considerable research gap because several TV series, such as *Sex Education*, feature non-binary characters and their representation can be challenging because of how gender is expressed across languages.

While in notional gender languages such as English only pronouns (*he/she/it*) and a few nouns, generally related to kinship (e.g. *mother/father*) or compounds and/or professions (e.g. *chairman/chairwoman*) are gendered, grammatical gender languages such as German also require extensive gender marking in other word classes, e.g. articles and adjectives (McConnell-Ginet, 2013; Stahlberg, Braun, Irmen, & Sczesny, 2007). For this reason, gender-fair language (GFL) strategies proposed to avoid masculine generics and/or to represent genders beyond the binary differ within and across languages. Gender-fair is used here to

subsume two approaches, i.e., gender-neutral and gender-inclusive (Sczesny, Formanowicz, & Moser, 2016). The former conceals gender, amongst others, by restructuring sentences using passive constructions, indefinite pronouns, and gender-neutral alternatives to common nouns. The latter makes all genders visible by utilizing neopronouns (e.g. *hen* in Swedish), neomorphemes (e.g. *e* in Spanish), typographical characters (e.g. * in German) or symbols (e.g. \varnothing in Italian). Human and machine translation of GFL is complicated because there is no one-fits-all solution to non-binary representation and some strategies may not be very common.

Current research in translation studies generally focuses on the use of GFL, or lack thereof, in audiovisual products or news reports translated from English into other languages such as Spanish, French, and German (Attig, 2022; Lardelli & Gromann, 2023c; López, 2022). Findings show that non-binary identities are usually erased or misgendered, thus translations fail to convey the original meaning of the source texts. Research in the field of natural language processing is starting to concentrate on gender-neutral MT (Piergentili et al. 2023). While this represents a first and important step towards fairer MT, gender-neutral language might not be desirable in reference to specific people whose gender is known, as in the case of non-binary people. Cross-fertilization among Translation Studies (TS) and Computational Linguistics (CL) is therefore of utmost importance and research on GFL ought to include different interest groups, e.g. non-binary people, language professionals, and MT experts (Gromann et al., 2023).

With this objective in mind, I conducted a first study on gender-fair post-editing inspired by Translation Process Research (TPR). Six professional translators were recruited and were asked to manually post-edit three English-to-German machine translations about non-binary actors joining the cast of different TV series. For each text, participants were instructed to utilize a different gender-fair language approach, i.e., gender-neutral rewording, gender-inclusive characters, and neosystems. Participants joined a videoconference, and their screens were recorded during the whole post-editing process. Subsequently, they took part in cued retrospective interviews. This setup made it possible to concentrate on the post-editing times, the screen activity tracking data, and the participants' experiences and impressions of the task.

While findings from the post-editing times and the retrospective interviews are the object of another publication (Lardelli & Gromann, 2023b), the focus of the present paper is on bias in the MT outputs and the screen activity tracking data to reconstruct the GFL post-editing process. Screen recordings were used to produce observation protocols, and different screen activities, i.e., broken words, changes, searches, and pauses, were analyzed to investigate the cognitive processes of the gender-fair post-editing process. Findings from the analysis of bias in machine translations show that MT systematically misgenders non-binary people and interprets singular *they* as plural. The analysis of the screen recordings suggests substantial differences among GFL approaches in screen activity, with high variability among participants. While the number of changes due to bias in the MT outputs is generally high among assignments, the number of searches dramatically increased when participants were using neosystems. The results from this case study can contribute to raising awareness of the strengths and limitations of different GFL approaches while considering the cross-linguistics expertise of translation professionals. I also support Attig's (2022) call for community-informed translation – a concept that should also be extended to (gender-fair) machine translation research.

2 Preliminaries

To provide a basis for gender-fair post-editing, this section introduces the interplay between gender and language as well as GFL approaches in English and German.

The term gender has multiple meanings. Before the introduction in the 1950s of the concept of gender roles, i.e., social behavior and characteristics that are associated with masculinity and femininity, the term gender was used to refer to grammatical gender (Haig, 2004: 89). Grammatical gender is a linguistic feature of nouns which pertains to their classification, e.g. as masculine, feminine, and neuter (Corbett, 1991; Hockett, 1958). Languages can be classified, on the basis of their gender systems, into (i) grammatical gender, (ii) notional gender, and (iii) genderless (Corbett, 1991; Stahlberg et al., 2007). In (i), such as German, each noun has a gender, i.e. including those referring to inanimate objects and abstract concepts. Other word classes, e.g. articles and pronouns, must be inflected accordingly. In (ii), such as English, gender is distinguished through pronouns, (*he/she/it*), and most nouns except for those referring to kinship (*mother/father*), professions and/or compounds (*chairman/-woman*) are gender-neutral. Finally, in (iii), such as Turkish, there is generally no gender distinction.

From the second half of the 1960s, the term gender has been used by feminists to refer to differences in societal roles and opportunities of men and women that are the results of social processes and cannot be ascribed to biology (Von Flotow, 1997: 5). Nowadays, gender is defined as a biopsychological construct that conflates biology (e.g. hormone levels), psychology (e.g. the sense of one's gender), and society (e.g. gender expectations, roles, and norms) (Barker, 2018: 99f). Gender is not binary and there is a wealth of identities beyond man and woman that may be subsumed under the term non-binary.

Because in grammatical and notional gender languages there is an overlap between grammatical and referential gender, i.e., the extralinguistic reality, linguistic strategies have been proposed to overcome the binary distinction between men and women. For instance, in English *they* has gained usage as a gender-fair singular pronoun (Baron, 2020) and gender-neutral alternatives to nouns have been introduced (e.g. *chairperson* instead of *chairman/-woman*) (APA Style, 2019). On the contrary, in German, there are four approaches (see Hornscheidt, 2012; Hornscheidt & Sammla, 2021; Lardelli & Gromann, 2023c for more in-depth overviews):

- **Gender-neutral rewording:** reference to a person's gender is avoided by means of passive constructions, indefinite pronouns, and gender-neutral terms amongst others. A popular form in German is the use of participial forms as alternatives to gendered nouns in the plural (e.g. *die Lesenden* (EN: the readers)). These forms are partially common both in the written and oral language.
- **Gender-inclusive characters:** masculine nouns are made feminine by adding the suffix *-in*, e.g. *Leser* (EN: masculine reader) becomes *Leserin* (EN: feminine reader). For this reason, gender-inclusive forms are created by separating the feminine suffix with different characters such as gender star (*) or underscore (_) as in *Leser*in* (masculine*feminine noun). In pronouns and articles, characters can be concatenated differently, e.g. *der*die* (masculine*feminine article), *die*der* (feminine*masculine article), or *di*er* (blended articles). There is no acknowledged norm or standard for doing so and gender-inclusive forms are relatively common in written language, especially to avoid masculine generics.
- **Gender-neutral characters:** characters such as *x* are used to replace gender suffixes as in *dix Lesex* (gender-neutral article and noun). This strategy is mostly used in activist circles when the gender identity of a person is not known or irrelevant to the context of the conversation.

- **Neosystems:** this strategy consists of the introduction of a fourth gender in the German language along with masculine, feminine and neuter. A new set of pronouns, suffixes, and grammar rules are hence proposed as in *din Lesernin* (non-binary article and noun). While neosystems are not very common, they are devised for and by non-binary people.

The example of English and German shows that GFL strategies differ within and across languages. Such (cross-)linguistic differences complicate the selection of a strategy for both human and machine translation.

3 Method

While the proposed method is discussed at greater length in Lardelli & Gromann (2023b), a brief description is provided in this section focusing on the analysis of the bias in MT outputs as well as the observational notes and screen recordings used to investigate the screen activity required to integrate GFL in post-editing.

The study was inspired by Translation Process Research (TPR) (Jakobsen, 2017) and, especially, Albl-Mikasa et al. (2017). It combines non-participant observation, screen recordings, retrospective interviews, and translation annotation. Six professional translators with at least three years of practical experience were recruited. Prior to the study, they received a translation brief with instructions on the tasks, post-editing guidelines from the Translation Automation User Society (TAUS),¹ and a handout on various strategies for gender-fair language to prepare for their participation.² They also compiled a survey to collect data on their profiles, including work experience and use of GFL.

As shown in Table 1, participants received three different texts of approximately 150 words on three English language TV series, i.e., *Sex Education*, *Grey's Anatomy*, and *Sort Of*. The texts discussed non-binary actors joining the cast of said series and were retrieved from TV news websites. They were translated with DeepL in July 2022. Participants received a text file with a table containing the source texts with the machine translations. They were asked to manually post-edit the texts using different GFL approaches for each one, i.e., (i) gender-neutral rewording, (ii) gender-inclusive characters, and (iii) neosystems. Participants could freely select specific strategies from the handout provided, e.g. gender star (*) amongst others for (ii).

¹ <https://info.taus.net/mt-post-editing-guidelines>

² The instructions and texts of this case study are available at <https://doi.org/10.5281/zenodo.789832>

Text No.	TV Series	Instructed Gender-Fair Approach	Word Count	No. of Gendered Phrases
1	Sex Education	Gender-Neutral Rewording	152	9
2	Grey's Anatomy	Gender-Inclusive Characters	151	12
3	Sort Of	Neosystems	163	10

Table 2. Details on the study materials

The texts contained references to both non-binary actors as well as mixed-gender groups. German was selected as the target language because, as a grammatical gender language, it needs extensive gender marking as compared to English. Moreover, since MT is known to be best equipped for high-resource languages due to the amount of training data available (Forcada, 2017), English is a good selection as the source language.

The study was conducted online as it aimed for the most authentic and unintrusive experimental setting. Participants joined a videoconference that was open in the background while they shared their screens, which were recorded. They could work in their familiar environment and were instructed to work under usual conditions. Nevertheless, they were required to use just one screen for the whole post-editing process to be recorded. Subsequently, they were interviewed about their impressions, strategies, and experience with GFL within the context of the study.

Similarly to other post-editing research (e.g. Carl et al. 2015), cognitive processes were investigated by analyzing post-editing times, screen activity tracking data, and participants' retrospective accounts. The focus of this paper is on the bias in the MT outputs and the screen activity. For each text, respectively 9, 10, and 12 phrases containing gender references to be post-edited were selected and pasted into an Excel sheet where they were annotated for three types of bias: (i) misgendering, (ii) mistranslations of singular *they*, (iii) masculine generics. The author's observational notes and the screen recordings of the post-editing process were used to produce observational protocols. The source texts and their machine translations were divided into segments in an Excel sheet. For each segment, the number of screen activities was entered, along with a description. Screen activities analyzed included: (i) broken units,³ (ii) changes,⁴ (iii) searches, (iv) pauses.⁵ In addition, more general comments on the GFL post-editing process were included in the protocols. Due to space constraints, Table 2 shows an example of one change made by P5 in the third segment of the first text along with one comment on the post-editing process. The total number of screen activities per assignment, segment, and participant as well as the median were calculated.

³ Broken units refer to chunks of text left unfinished and followed by, e.g., a pause or a search.

⁴ Text modifications such as insertions and deletions, typos excluded.

⁵ Pauses were measured starting from three seconds since the analysis of the screen recordings was conducted manually.

MT	Changes		Comments
	N.	Type	
Die Hauptrolle spielt der nicht-binäre Schauspieler und Musiker Dua Saleh als Cal	6	(1) "Die Hauptrolle spielt der nicht-binäre Schauspieler" ("der" deleted). ...	The participant takes some pauses but does not actually post-edit – they just focus on different gendered references and then highlights them

Table 3. Extract from an observational protocol (P5)

4 Results

First, the results of the bias analysis in the machine-translated texts are presented. Subsequently, the participants' profiles are briefly presented before comparing the screen activity tracking data among assignments and participants. Finally, insights into challenges encountered in each assignment are described.

4.1 Gender Bias in Machine Translation

As mentioned in Section 4, phrases containing gender references to be post-edited were annotated for three types of gender bias, namely (i) misgendering, (ii) mistranslations of singular *they*, (iii) masculine generics. Table 3 provides an overview of the results of the gender bias analysis for each machine translation. Note that one phrase could contain different types of bias, thus the number of gender-biased phrases and the total number of annotations in Table 3 do not always match.

Text No.	No. of Gender-Biased Phrases	No. and Typology of Gender Bias Annotation
1	6/9	Misgendering (2), mistranslations of singular <i>they</i> (3), masculine generics (1)
2	10/12	Misgendering (5), mistranslations of singular <i>they</i> (4), masculine generics (2)
3	9/10	Misgendering (6), mistranslations of singular <i>they</i> (3), masculine generics (1)

Table 4. Overview of the gender bias in the machine translations

Gender bias was found in almost every phrase analyzed. The non-binary actors and characters mentioned in the texts were systematically misgendered using masculine forms. In one case only, DeepL misgendered the non-binary person mentioned using a female form, i.e., the noun

“neuroscientist” was translated into “Neurowissenschaftlerin” (EN: feminine neuroscientist) in the second text. Plural references were also translated into the masculine. This led to a double annotation in the third text: non-binary celebrities Baig and Fabo are referred to as “co-creators” of the Canadian TV series *Sort Of*. The German machine translation “Co-Autoren” can be thus interpreted as an instance of misgendering and a masculine generic. Finally, singular *they* was consistently translated into the third-person plural pronoun “sie” in German.

During the annotation of the third text, two MT errors were found that deserve further attention. First, there was a co-reference mistake. In the English language text, “audiences” are mentioned in the first sentence, and “they” is used in the second as an anaphoric pronoun. In the German machine translation, the plural noun was translated into the singular, i.e., “das Publikum” (EN: the public), and the third-person singular pronoun “sie” was used as an anaphor instead of the grammatically correct “es” (EN: it). Second, the series is about a nanny who is genderfluid. This identity term was translated into “geschlechtsspezifisch” (EN: gender-specific).

4.2 Participant Profiles

Table 4 provides an overview of the participant profiles, which are rather heterogeneous. Participants were aged between 25 and 59 years. Four were women, while two were men. They all worked as professional translators and/or interpreters and had three to 20+ years of experience. Each participant was also familiar with post-editing and GFL, although with different degrees of experience. All used GFL in their daily work with the exception of P2, who was a patent translator and stated that GFL is not welcome in this field. While two participants indicated that they use gender-neutral rewording or gender-inclusive characters depending on the client and/or assignment, three only use the latter strategy. Participants were also asked to rate the difficulty of GFL, and their answers were generally positive to neutral. Only P6 indicated that GFL was difficult.

Participant	Age	Gender	Work Exp.	PE Exp.	GFL Exp.	GFL Use	GFL Difficulty
P1	32-38	Man	6-10	Extensive	Yes	Depends on client/assignment	Easy
P2	39-45	Woman	16-20	Extensive	Little	No	-
P3	53-59	Woman	20+	Little	Little	Gender-inclusive characters	Neither difficult nor easy
P4	32-38	Woman	6-10	Extensive	Yes	Depends on client/assignment	Really easy

P5	25-31	Woman	3-5	Little	Little	Gender-inclusive characters	Easy
P6	39-45	Man	11-15	Extensive	Yes	Gender-inclusive characters	Difficult

Table 4. Participants' profiles

4.3 Screen Activity Tracking

Figure 1 presents an overview of the median occurrence for each screen activity, i.e., broken units, changes, searches, and pauses, as well as their sum in each assignment, namely gender-neutral rewording, gender-inclusive characters, and neosystems. The median is used here due to the considerable differences found in the number of screen activities among participants which would have led to a skewed mean. Screen activity tracking was highest in the third assignment requiring neosystems, and lowest in the first assignment requiring gender-neutral rewording. The number of changes was high across assignments since the MT outputs were considerably biased and required extensive post-editing to become gender fair. The number of searches dramatically increased in the third assignment.

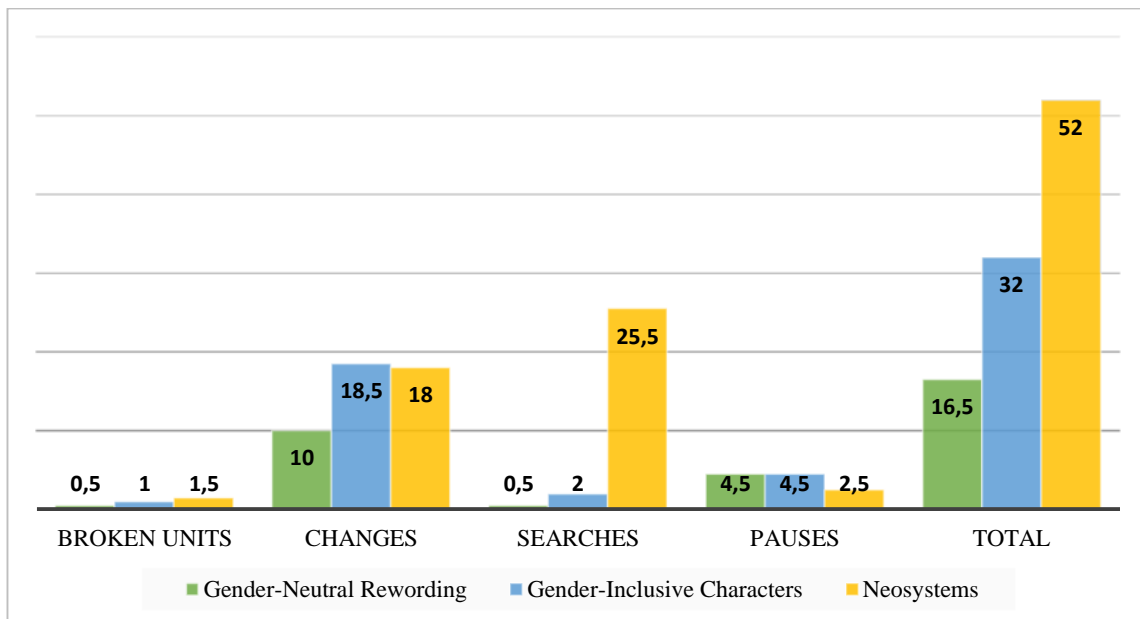


Figure 3. Overview of screen activities for each assignment

While the median for the occurrence of broken units was very low and similar across assignments, respectively 0.5, 1, and 1.5, the number of changes was lower in the first assignment (10), and increased in the others with a median of 18.5 occurrences for the second and 18 for the third assignment. Participants made very few searches when rewording (Mdn=0.5) and using gender-inclusive characters (Mdn=2) but this increased considerably when using neosystems (Mdn=25.5). Two participants, namely P2 and P5, even made respectively 72 and 71 searches. P3 read aloud some of the GFL handout passages provided

before the study in an attempt to concentrate and understand how the selected neosystem worked. The constant need for resources was common to all translators in the third assignment.

During this assignment, four participants, i.e., P1, P3, P5, and P6, split their screens side by side to place the machine translation next to the GFL handout. This allowed them to edit gendered references without having to constantly open a new window. Moreover, P3, P4, and P5 allocated all of their mental resources to GFL and edited all the gender references in the third text first. Only subsequently did they proceed to post-edit the whole text, e.g. adapting the style to specifications included in the translation brief.

The searches conducted during the study can be divided into two different typologies. Participants actively looked for GFL strategies and how to use them. This also included using online dictionaries to find possible translations of a certain term or gender-neutral synonyms/alternatives. Furthermore, translators also looked for more information concerning the TV series mentioned in the texts. While doing so, they often focused on the descriptions of the non-binary characters and looked for pictures of them, as shown in Figure 2 which is a screenshot of P3's desktop.

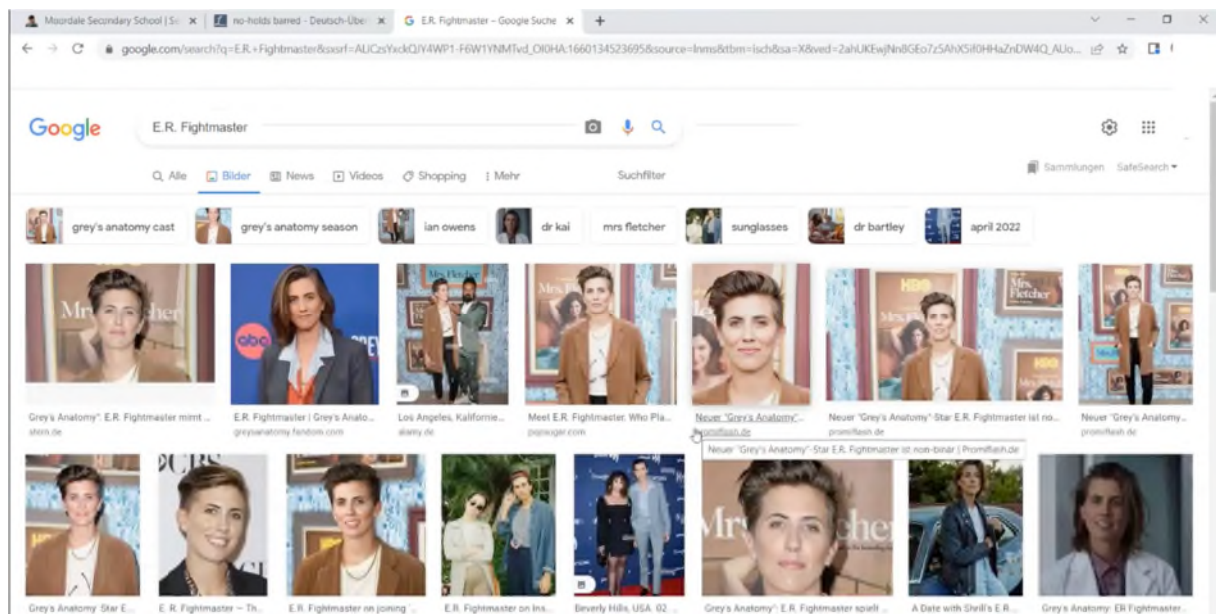


Figure 4. Image search to visualize the non-binary actor mentioned in one text

Finally, the number of pauses was the same in the first and second assignments (Mdn=4.5) and decreased in the third (Mdn=2.5). The screen activity tracking data of each participant are reported in Table 5.

BU indicates broken units, C changes, S searches, P pauses, and T the sum of the analyzed screen activities. P1 was the participant for whom the fewest screen activities were observed in each assignment, i.e., respectively 9, 12, and 18. P1 has extensive knowledge of GFL and is specialized in queer/feminist translation. P2 was the participant for whom the highest screen activity was observed. Throughout the assignments, P2 was one of the translators who made the highest number of changes and, especially, searches. P2 also made numerous pauses in the first assignment because very unsure about how to reword gender references. P2 was the only participant who did not use GFL in everyday work. The table also shows that the highest screen activity was observed in the third assignment for all participants, while the lowest was in the

first text for four out of the six translators. In the following paragraphs, I will concentrate on specific challenges encountered in each assignment to show the potential strengths and limitations of each GFL approach.

Participant	1st Assignment					2nd Assignment					3rd Assignment				
	<i>B</i> <i>U</i>	<i>C</i>	<i>S</i>	<i>P</i>	<i>T</i>	<i>B</i> <i>U</i>	<i>C</i>	<i>S</i>	<i>P</i>	<i>T</i>	<i>B</i> <i>U</i>	<i>C</i>	<i>S</i>	<i>P</i>	<i>T</i>
P1	0	8	0	1	9	0	1 2	0	0	1 2	1	5	1 1	7	1 8
P2	4	1 9	7	2 3	5 3	1	2 1	2 1	8	5 1	4	2 7	7 2	9	1 12
P3	0	7	0	3	1 0	2	2 5	3	7	3 7	0	1 6	2 0	1 2	4 8
P4	0	7	4	6	1 7	1	1 5	1 6	5	3 7	3	2 0	7 1	1	9 5
P5	2	1 4	1	1 1	2 8	2	2 0	1	4	2 7	4	1 8	3 1	3	5 6
P6	1	1 2	0	3	1 6	1	1 7	0	2	2 0	0	1 9	7	2	2 8

Table 5. Overview of the number of screen activities for each participant across assignments

Gender-neutral rewording makes it possible to avoid gender-specific references using the existing language inventory. However, alternatives to common nouns must be found. For instance, when post-editing the German translation for “actor and musician” participants needed to build compounds with gender-neutral terms such as “Star” or “Talent”. An extract from P5’s screen recording is transcribed here as an example:

pause: 10 s | decides to read and post-edit other segments of the machine translation | pause: 25s | decides to read and post-edit other segments of the machine translation | pause: 37s | decides to read and post-edit other segments of the machine translation | pause: 25s | change: ~~der~~ nicht-binäre Schauspieler (EN: the non-binary actor, masculine used in the German MT) | search: gender-neutral for actor | change: ~~der nicht-binäre Schauspieler~~ eine nicht-binäre schauspielerisch und musikalisch tätige Person (EN: a non-binary person active in cinema and music) | pause: 4s

Similarly, there is no direct gender-neutral alternative for the term “student” when this refers to high school. Participants generally opted for terms such as “newcomer” or “new entry” in German, as P2:

pause: 13s | search: synonym | change: SchülerNeuzugang (EN: new entry) | pause: 6s | change: **Klassen**Neuzugang (EN: new entry in the class)

Finally, since there is no equivalent for singular *they* in German, translators avoided pronouns by repeating the proper name or rewording whole text passages. P4's workflow is reported here as an example:

pause: 15s | change: SieCal (i.e., the name of the person mentioned in the text) | search(x2):
handout | pause 28s

Gender-inclusive characters are relatively common and straightforward to use. However, there are some exceptions as in the case of “doctor”, i.e., a noun whose word stem differs in the masculine (Arzt) and feminine form (Ärztin). This can be exemplified by reporting on P3's workflow:

pause: 13s | C: ArztMediziner*in (EN: physician) | pause: 6s | change: Mediziner*inMitglied
in das Ärzteteam (EN: member of the doctor team) | pause:15s

Moreover, there is no standard in the order and/or concatenation of masculine and feminine forms when using gender-inclusive characters. This may create confusion, as in the case of P2:

search: handout | change: di*er von dem nicht-binären Schauspieler (EN: who (is played by) the non-binary actor) | search: use of gender star + handout (the participant found that the form they had just used is rather uncommon) | change: die*der von dem nicht-binären Schauspieler

In the third assignment with neosystems, a challenge for the post-editing process was represented by the term “nanny” because the profession is stereotypically associated with women and there are no masculine equivalents. This caused great difficulties, for instance, for P2 who made 34 searches, reading the provided handout on GFL or looking for alternative solutions on the internet. Finally, participants struggled to understand how the selected neosystems should be used, especially in the case of pronoun and adjective declensions as exemplified by P4's workflow:

search: 9x handout (the participant actively looked for the pronoun “sin” which does not actually exist) + 7x online searches | changes: um etwas mehr als ihrsinseinemseinimnimser offensichtlich unbefriedigendes Leben (here, the participant selects three different possessive pronouns before finding the correct one, i.e., “nimser” for the English source text “in pursuit of something more than their clearly unsatisfying life”)

5 Discussion & Conclusion

The results from the gender bias analysis confirm that current MT systems do not recognize non-binary pronouns and erase non-binary identities in their outputs. This is alarming because people are often exposed to machine-translated texts without being aware of this and gender bias against non-binary people is potentially infinitely propagated. Moreover, results show that identity terms referring to queer people are incorrectly translated, which represents a potential direction for future research on gender bias. Finally, the co-reference mistake found confirms the limitations of context-unaware MT systems.

The results from the post-editing study show differences in screen activity tracking both among assignments and participants. While gender-neutral rewording required less screen activity, this increased in gender-neutral characters and, especially, in neosystems. The number of changes was generally high, which confirms that the MT outputs are severely biased. The number of changes was higher in the second and third assignments than in the first. The reason might be that gender-neutral rewording is not a disruptive strategy that introduces new characters or gender suffixes in language, and one uses the existing language inventory. Although relatively common, gender-inclusive characters can be applied differently and this sometimes leads to uncertainties in how to concatenate masculine and feminine forms. On the contrary, neosystems are generally unknown and more time might be needed to understand how these new forms are used. This led not only to a high number of changes but also and especially searches. The number of searches was limited in the other two assignments. The reason for this might be that the first two approaches are relatively common in German-speaking countries, especially to avoid masculine generics. For instance, Austria is one of the few countries where the use of (binary) gender-fair language is mandatory in job advertisements and public administration (Feigl, 2009). During the post-editing process, participants generally looked online for more information concerning the TV series and/or the non-binary actors and characters mentioned in the assignments. Some participants also looked for pictures of the said people. It is important here to underline that gender identity does not equal gender expression and that one should not infer the gender identity of a specific person based on their physical aspect.

Insights into the screen activity tracking data reveal some challenges of each GFL strategy. Gender-neutral rewording has the advantage of using the existing language inventory without the need to introduce new pronouns and suffixes. However, it might be challenging to find gender-neutral alternatives to common terms or avoid personal pronouns. Gender-inclusive characters are a relatively straightforward inclusive approach. Nevertheless, the lack of a norm or standard leads to inconsistent use of the characters. Neosystems are generally proposed by non-binary people and are therefore the most appropriate solution to non-binary representation. Unfortunately, they are rather uncommon and translators might need time and training to learn how to use them correctly.

Attig (2022) calls for a community-informed translation practice for texts referring to or mentioning queer people. While this is of utmost importance, the (cross-)linguistic expertise of professional translators is also needed to account for the wealth of factors that play a role in selecting a specific gender-fair language strategy. These factors include, amongst others, client specifications, target culture, target public and accessibility. For this reason, cross-fertilization between disciplines such as translation studies and computational linguistics is needed to tackle gender bias beyond the binary. Participatory research approaches as in Burtscher et al. (2022) and Gromann et al. (2023) represent necessary steps towards more community-informed and gender-fairer machine translation. Researchers must consider that there is no one-fits-all

solution to binary representations and people have the right to (linguistic) self-determination. Furthermore, terminology related to identity terms and gender-fair language strategies constantly evolve over time.

In this first gender-fair post-editing study, I asked professional translators to post-edit three English-to-German machine translations on non-binary actors appearing in TV series by applying different gender-fair language strategies, i.e., gender-neutral rewording, gender-inclusive characters, and neosystems. The focus of the present paper was on the gender-biased MT outputs and the screen activity needed to produce gender-fair texts. Differences among assignments were found, with the fewest screen activities in the first assignment and the most in the third. Notably, participants conducted many searches to correctly utilize neosystems and the number of changes was high across assignments due to the biased outputs. The analysis of the machine translations shows an overuse of masculine forms which leads to the erasure of the non-binary people mentioned in the texts. Identity terms are also wrongly translated sometimes. Further research on the comparison of gender-fair language in translation and post-editing across different natural languages as well as the reception of gender-fair texts, especially with a focus on their accessibility, would be interesting.

Acknowledgements

I would like to warmly thank all the study participants for their motivation and interesting insights into the topic as well as UNIVERSITAS Austria for the financial and logistic support. I would also like to thank all the researchers I discussed my work with and whose insights helped improve this paper and my research in general.

References

- Albl-Mikasa, Michaela, Giovanna Fontana, Laura Maria Fuchs, Lena Meret Stüdeli, and Aline Zaugg. (2017). Professional translations of non-native English: ‘before and after’ texts from the European Parliament’s Editing Unit. *The Translator*, 23(4): 371–387. <https://doi.org/10.1080/13556509.2017.1385940>
- APA Style. (2019). *Gender*. <https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/gender> [last accessed 09.11.2023].
- Attig, Remy. 2022. *A call for community-informed translation: Respecting Queer self-determination across linguistic lines*. <https://doi.org/10.1075/tis.21001.att>
- Barker, Meg John. (2018). *Rewriting the rules: An Anti Self-help Guide to Love, Sex and Relationships*. Routledge.
- Baron, Dennis. 2020. *What’s Your Pronoun?: Beyond He and She*. Liveright Publishing.
- Burtscher, Sabrina, Katta Spiel, Daniel Lukas Klausner, Manuel Lardelli, and Dagmar Gromann. 2022. “Es geht um Respekt, nicht um Technologie”: Erkenntnisse aus einem Interessensgruppen-übergreifenden Workshop zu genderfairer Sprache und Sprachtechnologie. In *Mensch Und Computer 2022*, pages 106-118.
- Carl, Michael, Silke Gutermuth, and Silvia Hansen-Schirra. 2015. Post-editing Machine Translation. Efficiency, Strategies, and Revision Processes in Professional Translation Settings. In Aline Ferreira & John W Schwieter (Eds.), *Psycholinguistic and Cognitive*

- Inquiries into Translation and Interpreting*. John Benjamins Publishing Company, pages 145-174.
- Corbett, Greville G. 1991. *Gender*. Cambridge University Press.
- Feigl, Susanne. 2009. *Geschlechtergerechte Stellenausschreibung: unabhängiger Bericht der Gleichbehandlungsanwaltschaft i.S. § 3 Abs. 5 GBK-GAW-Gesetz*. Bundeskanzleramt Österreich, Gleichbehandlungsanwaltschaft bei der Bundesministerin für Frauen und Öffentlichen Dienst.
- Forcada, Mikel L. 2017. Making sense of neural machine translation. *Translation Spaces*, 6(2): 291–309.
- Gromann, Dagmar, Manuel Lardelli, Katta Spiel, Sabrina Burtscher, Lukas Daniel Klausner, Arthur Mettinger, ... Katharina Bühn. 2023. Participatory Research as a Path to Community-Informed, Gender-Fair Machine Translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, 49–59.
- Haig, David. 2004. The Inexorable Rise of Gender and the Decline of Sex: Social Change in Academic Titles, 1945–2001. *Archives of Sexual Behavior*, 33(2): 87–96.
- Hockett, Charles Francis. 1958. *A Course in Modern Linguistics*. Macmillan.
- Hornscheidt, Lann. 2012. *Feministische W_orte: ein Lern-, Denk- und Handlungsbuch zu Sprache und Diskriminierung, Gender Studies und feministischer Linguistik* (1. Aufl). Brandes & Apsel.
- Hornscheidt, Lann, and Ja'n Sammla. 2021. *Wie schreibe ich divers? Wie spreche ich gendergerecht? ein Praxis-Handbuch zu Gender und Sprache*. w_orten & meer.
- Jakobsen, Arnt Lykke. 2017. Translation Process Research. In John W. Schwieter & Aline Ferreira (Eds.), *The Handbook of Translation and Cognition*. Wiley Online Library, pages 19-49.
- Lardelli, Manuel, and Dagmar Gromann. 2023a. Gender-Fair (Machine) Translation. In *Proceedings of the New Trends in Translation and Technology Conference - NeTTT 2022*, pages 166–177.
- Lardelli, Manuel, and Dagmar Gromann. 2023b. Gender-Fair Post-Editing: A Case Study Beyond the Binary. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260.
- Lardelli, Manuel, and Dagmar Gromann. 2023c. Translating non-binary coming-out reports: Gender-fair language strategies and use in news articles. *The Journal of Specialised Translation*, 40: 213–240.
- López, Ártemis. 2022. Trans(de)letion: Audiovisual translations of gender identities for mainstream audiences. *Journal of Language and Sexuality*, 11(2): 217–239. <https://doi.org/10.1075/jls.20023.lop>
- McConnell-Ginet, Sally. 2013. ` Gender and its relation to sex: The myth of ‘natural’ gender. In Greville G. Corbett (Ed.), *The Expression of Gender*, 3–38. <https://doi.org/10.1515/9783110307337.3>
- Piergentili, Andrea, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. (2023). Gender Neutralization for an Inclusive Machine Translation: from Theoretical

Foundations to Open Challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9: 845–874. https://doi.org/10.1162/tacl_a_00401

Sczesny, Sabine, Magda Formanowicz, and Franziska Moser. 2016. Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00025>

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. In *Frontiers of Social Psychology. Social communication*. Psychology Press, pages 163-187.

Von Flotow, Luise. (1997). *Translation and Gender: Translating in The 'Era of Feminism'*. Routledge.

The use of speech technologies and machine translation in institutional translation practices

Justus Brockmann

University of Vienna

justus.brockmann@univie.ac.at

Alina Secară

University of Vienna

alina.secara@univie.ac.at

Dragoş Ciobanu

University of Vienna

dragos.ioan.ciobanu@univie.ac.at

Abstract

Technology is rapidly changing the way translators work, and institutional translation is no exception. We report on a survey investigating technology use in institutional translation practices, focusing on two types of technology: machine translation (MT) and speech technologies (automatic speech recognition (ASR) and text-to-speech synthesis (TTS)). The survey was answered by 93 respondents from different national and international institutions. We report on their use and perceptions of machine translation, their experience with voice-enhanced working practices, and their priorities for technology implementation and training. MT implementation is widespread in our sample, with 74% reporting having integrated MT. Translators' opinions about MT quality are mostly positive, which is significantly related to having access to a customised MT system. Responses suggest that there is room for debate on how to define MT-supported workflows, and that these should not lose sight of the human factor. The adoption of speech technologies is modest. Low use of ASR (2%) is linked to access to good quality MT, while TTS is gaining visibility both in terms of existing use (6%) and interest in receiving TTS training. In terms of implementation and training priorities, MT and post-editing lead the list of topics mentioned most frequently, followed by speech technologies.

1 Introduction

Translating institutions have for many years accounted for a sizeable share of the translation work carried out in the world (Koskinen, 2008, p. 2). As demand for translation rises in societies at large, those who provide translations increasingly rely on technology to meet this demand, and the institutional space is no exception. This is true for long-established types of technology, such as computer-assisted translation (CAT) tools, and machine translation (MT), which – despite the attention it has received for decades – is far from permeating all types of translation processes. More recent technological advancements in areas such as speech technologies (automatic speech-to-text or text-to-speech conversion) are also increasingly attracting attention from translation scholars and practitioners, warranting further research in this area.

We report on a survey designed to investigate the use and perceptions of technology in institutional translation practices, focusing in particular on MT and speech technologies. After providing an overview of related research in section 0, we outline how the survey was designed and distributed, as well as who our respondents were, in section 0. In section 0, we explore respondents'

- (1) use and perceptions of MT,
- (2) use and perceptions of speech technologies, and
- (3) priorities regarding technology training and implementation.

Limitations of the survey and our conclusions are discussed in sections 0 and 0, respectively.

2 Background

Societies create institutions to serve a variety of needs (Koskinen, 2011), and wherever the work of these institutions crosses linguistic or cultural borders, they are likely to rely on translation (Pym, 2008). The “vast and variegated field” (Martín Ruano, 2019, p. 269) of institutional translation has been traditionally considered a *missing factor in translation theory* (Mossop, 1988; see also Kang, 2014), but interest in the institutional field has grown considerably in Translation Studies over recent decades (Svoboda et al., 2022). Translation carried out for the European Union (EU) institutions in particular has attracted the attention of scholars (e.g., Svoboda et al., 2017). Yet, translation by and for other institutions has seen less coverage (Kang, 2019).

We use Koskinen’s (2011) definition of institutional translation to refer to translation carried out for a concrete organisation that has been assigned by society with some sort of control or governance function, and where translation is typically collective, anonymous and standardised. The organisations we include in our definition all fulfil some sort of governmental function, and they rely on translation, at least to some extent, to communicate with the societies they serve. More specifically, we include in this definition supranational and intergovernmental organisations such as bodies of the EU and the United Nations (UN), as well as national and regional institutions. As an aside, we also think that Koskinen’s idea of the main features of translation in such organisations needs to be updated to account for current practices. Trends such as increasing outsourcing, decreasing capacity for full in-house revision rather than spot-checks, and a growing reliance on artificial intelligence (AI) tools have already changed what ‘collective’ and ‘standardised’ meant in several such institutions before 2011.

The ongoing technologisation of translation is rapidly changing the institutional field (Lafeber, 2022), and several studies have addressed the (non-)adoption and use of technology in individual institutions (Cadwell et al., 2016; Macken et al., 2020; Rossi and Chevrot, 2019; Vardaro et al., 2019). Surveys targeted at institutional translators have covered technology-related aspects such as translator training (Svoboda and Sosoni, 2022), skills and competences relevant for institutional translators (Froeliger et al., 2022; Lafeber, 2012, 2022), and translators’ perceptions of MT (Rossi and Chevrot, 2019). These surveys, among other findings, highlight how the profile of the institutional translator is changing, and how technology plays an increasingly important role when it comes to navigating the pressures of decreasing turnaround times while at the same time retaining the value of human work which cannot be replaced by any automation solution available today.

Other surveys targeted at professional translators and language service providers in general have frequently addressed the use of MT. While almost a decade ago, around 40% of respondents to these surveys were using MT (Gaspari et al., 2015; Zaretskaya et al., 2015), more recent surveys report higher uptake of between 70% and 75% (ELIA et al., 2023; Farrell, 2023). This underlines the growing importance of MT in translation processes – but it also shows that MT is far from being used by all professional translators.

Apart from MT, there is also a growing research interest in the application of speech technologies in translation and related tasks, which can also be attributed to the potential of these technologies to address some of the issues brought about by integrating MT into current workflows. We use speech technologies as an umbrella term that encompasses both automatic

speech recognition (ASR), which automatically converts human speech input from a microphone or audio file into written text, and text-to-speech synthesis (TTS), which automatically converts written text into synthetic speech audio.

Research into ASR has already indicated its potential benefits and use cases in translation tasks (Ciobanu, 2014, 2016; Dragsted et al., 2011; Mesa-Lao, 2014; Zapata et al., 2017), which include improved working speed and translator ergonomics. These advantages have also seen some coverage by industry news outlets, e.g. in a 2022 *Slator*¹ article. When it comes to actual industry use, the 2023 European Language Industry Survey, ELIS (ELIA et al., 2023), shows that, on the part of independent translators, the use of ASR is still modest. However, slightly more of them started using ASR in 2022 compared to 2021, and a growing percentage would like to invest in ASR software in 2023 (ibid., Fig. 87). Chereji (2024) also found modest use of ASR in her survey of medical translators, where 15% of respondents indicated having used the technology. The annual *Nimdzi Language Technology Atlas* has been featuring ASR for years among the key technologies used in the industry (Nimdzi Insights, 2023), although mainly focusing on solutions for automatic transcription, captioning, and subtitling.

TTS, on the other hand, has so far received far less scholarly attention than ASR, but there is initial research pointing to the benefits and implications of implementing TTS in revision (Ciobanu et al., 2019) and post-editing (Wiesinger et al., 2022; Rios Gaona et al., 2024) workflows.

However, to date investigations into contemporary applications of speech technologies in institutional translation have been scarce. The latest ELIS surveys (e.g., ELIA et al., 2023) identify some degree of implementation of ASR by public administrations for subtitling and dubbing, but some language professionals in the EU institutions, for instance, are also known to use ASR to dictate text into their computer instead of typing (European Parliament et al., 2019). Liyanapathirana et al. (2019) surveyed translators at several international organisations on their current use of ASR and their openness towards using it for post-editing. They report that a surprisingly high number of respondents are already using the technology when translating from scratch (9 in 17) and two respondents even use it for post-editing. In conference interpreting – a field that is also seeing rapid technological change, significantly catalysed by the COVID-19 pandemic – speech recognition is discussed among the major technologies that have the potential to support interpreters in the booth through automatic transcription (Jayes, 2023). This is also reflected in a recent report on the effects of technologies on the work and roles of translators and interpreters (Orlando et al., 2024). It highlights the rise in popularity of speech tools, especially linked to interpreting practices, but points to the relatively scant amount of research in this area. To our knowledge, there is no published research yet on institutional translators' interaction with TTS tools.

The present survey is part of a wider research interest in the use of technology in the institutional translation context as part of the first author's PhD project. Moreover, we are building on our previous research into voice-enhanced practices in revision (Ciobanu et al., 2019), post-editing, and translation processes (Brockmann et al., 2022; Ciobanu, 2016; Wiesinger et al., 2022), as well as their adoption both in academia and the industry (Ciobanu, 2014; Zapata et al., 2023).

¹ <https://slator.com/how-translators-and-post-editors-benefit-from-speech-technologies/>

3 The survey

3.1 Design

The survey was set up in English, since it was assumed that this language would make it accessible to a large share of the targeted population. It featured 34 questions, grouped into four main blocks:

- (1) the respondents' professional backgrounds,
- (2) their use of MT,
- (3) their experience with voice-enhanced working practices, and
- (4) their priorities in terms of technology training and implementation.

Both closed-ended and open-ended questions were used to elicit standardised quantitative data on the one hand, while also giving respondents the opportunity to explain their answers to a previous question and, for the last block, to allow them to freely express their technology training and implementation priorities, as well as any general comments on the survey. Respondents' answers to the survey were anonymous in principle, but they were asked to specify the institution they work for. We provide access to the full survey questionnaire via a GitHub² repository.

3.2 Distribution

The Limesurvey³ tool was used to set up the survey as an online questionnaire, and distribute it using a hyperlink. The survey was open from November, 2022, to June, 2023. It was distributed through two different channels:

- (1) the Universities' Contact Group (UCG) of the International Annual Meeting on Language Arrangements, Documentation and Publications (IAMLADP), a network of representatives from over 80 international organisations employing conference and language service providers, and
- (2) the authors' contacts at intergovernmental organisations and national institutions.

A snowball sampling method was used. This meant that the persons contacted were asked to forward the survey to colleagues that fit the profile. As specified in the introduction text, the study was targeted at professionals involved in (written) translation projects carried out for an international organisation or a public institution (regional / national / intergovernmental). Before filling in the survey form, respondents gave written consent for their data to be collected, stored, and analysed for the purposes of the study.

As per Koskinen's (2011) definition above, there is an assumed standardised nature of institutional translation which can manifest itself in the institutions' use of *predictable* language-independent CAT tools, language-dependent and language-independent quality

² <https://github.com/HAITrans-lab/HAITrans-2023-institutional-survey>

³ LimeSurvey GmbH, Hamburg, Germany: <http://www.limesurvey.org>

assurance (QA) processes, or language-dependent style guides. These can be combined to a more or less similar effect for the majority of the language pairs the institutions handle. With the advent of technologies based on *unpredictable* neural networks which learn from unequal datasets (such as MT or speech technologies), however, linguistic parity is no longer possible, resulting in an interesting variation in the take-up and general institutional practices.

3.3 Respondents

Our 93 respondents represent a range of different institutions, most of them based in Europe, including large intergovernmental organisations like the bodies of the EU and the UN, as well as national administrative bodies and central banks. Numerous respondents work for the European Commission (44%). Among the national organisations, German institutions are by far the most frequent (23% of total responses).

The vast majority of respondents are in-house translators (75%). The other roles indicated include institution officials (9%), revisers (5%), and project managers (3%), among others. Most respondents have extensive experience working in their current roles and with their current employers, 82% having been employed for 5 years or more at their current institution.

Our respondents work in 80 unique language directions. English into German and German into English are the most prominent by far (34 and 27 respondents, respectively), followed by French into German (13), and Spanish into German (11).

When it comes to domains, *Administration, banking, business, economics and finance* and *Law* lead by a large margin, with 75% and 49% of respondents, respectively, translating in those domains. Further popular domains include *Agriculture and environment* (25%), *Social sciences* (17%), and *Industry* (15%). Our taxonomy of domains is based on Gaspari et al. (2015).

The linguistic activities our respondents typically engage in are, first and foremost, translation (96%) and revision (86%). Far fewer respondents said they typically perform post-editing (34%), editing (20%), and translating into plain language (11%), among others.

3.4 Statistics

Where appropriate, potential relationships between variables were tested for significance. Since all the variables measured through the closed-ended questions in our survey either have a nominal or ordinal level of measurement, Pearson's chi-squared test for independence was used. We report the chi-squared value (χ^2), the degrees of freedom (df) of the statistic, and the p value. However, the chi-squared test runs the danger of being inaccurate for contingency tables that contain multiple cell values lower than 5. We therefore computed Fisher's exact test whenever this was the case. All tests were run using the *R* language and environment for statistical computing (R Core Team, 2023) within the RStudio interface (Posit team, 2023). We used the conventional 95% confidence level for the significance tests.

4 Results and discussion

4.1 Use of machine translation

74% of respondents report that MT is integrated into workflows at their institutions (see Table 5). In a clear majority of these cases, MT is being used daily. Of the 24 respondents who do not have access to an MT solution implemented at their institution, 17 felt that their

organisation is likely or certain to start using MT in the future, while only 2 said this was unlikely to happen.

Yes, it is being used daily	54.84%
Yes, it is being used regularly	3.23%
Yes, it is being used occasionally	16.13%
No	25.81%

Table 5. Is machine translation integrated into workflows at your organisation / institution?

Most respondents use an in-house developed MT system (see Table 6). This is because all of them work either at the EU or the UN, and therefore have access to the in-house systems set up by these institutions. Another 11% of respondents from other institutions report using the EU's eTranslation system per the *Other* answer option. Freely available online MT is used by almost a quarter of respondents. This is likely due to its easy accessibility, but free online tools carry some well-documented risks (e.g., Canfora and Ottmann, 2020; DePalma, 2014) that professionals who have not received MT training are less likely to be aware of. It is therefore not surprising that those respondents who have received MT or post-editing-related training use free online MT significantly less than those who have not ($\chi^2 = 6.7155$, $df = 1$, $p = 0.01$). It is also interesting to note that 13 of our participants report using more than one type of system, which might be due to the fact that different systems have different strengths and weaknesses.

In-house developed machine translation system	44%
Freely available online machine translation system	23%
In-house machine translation system set up by external company	13%
eTranslation implemented by national institution (<i>Other</i>)	11%
Paid subscription to an external machine translation provider	10%
None	14%

Table 6. What type(s) of machine translation do you use in your work? (multiple response)

Opinions of MT quality are generally positive (see Table 7). Moreover, since we also asked respondents whether the MT system they are using is customised, we were able to identify a significant relationship between having access to a customised MT system and having a positive opinion about MT quality (Fisher's exact test: $p = 0.033$). Customised engines are predominantly found in international organisations, where 71% of respondents state they have a customised engine, compared with 24% of respondents from national institutions.

not usable	0%
poor quality	1%

medium quality	33%
good quality	49%
excellent quality	13%
no opinion	4%

Table 7. What is your opinion about the quality of machine translation currently available to you and / or your organisation / institution?

Interestingly, 16% of respondents who answered the question about MT customisation did not know whether the MT system available to them was customised, which also supports the call for emphasising MT-related training and increasing what has been termed *MT literacy* (Bowker and Ciro, 2019) among professional translators.

58% state that they currently post-edit MT output, which is a far higher share than those who included post-editing in their typical linguistic activities as mentioned at the end of section 0. However, it is still well below the three quarters of respondents working at institutions that have implemented MT. This can be explained to some extent by differences in opinion about what activities are encompassed by the term *post-editing*, as we will discuss below.

Among those currently involved in post-editing, experience differs: a small majority of respondents indicate 1-5 years of experience (52%). Post-editing is mostly carried out within CAT tools, but nevertheless, a small percentage (7%) of those who perform post-editing do so exclusively outside of CAT tools. In addition to post-editing, 65% of respondents report using MT for other purposes, such as gisting when receiving large source documents and needing to decide which parts represent a translation priority.

We also asked respondents who are not using MT in their translation work to specify their reasons for not using it (n = 39, see Table 5). 19 of those who do not use MT are working at an institution where there is no MT implemented, and 17 believe the MT quality is not satisfactory. 8 respondents report being faster when not using MT. Among the free-text reasons provided are MT's mishandling of regional-specific terminology, decreased creativity, and security. It is useful to note that 3 respondents specifically consider using MT suggestions offered through a translation memory (TM) not to constitute post-editing, because they see MT as being just one of the tools they use to meet the objective of delivering a translation. We will discuss this further below, as we believe this to constitute a central factor in the debate surrounding MT adoption.

My employer does not provide a machine translation solution	49%
The quality of machine translation output is not good enough	44%
I work faster without using machine translation	21%
Ethical reasons	10%

Financial reasons	3%
<i>Other</i>	28%

Table 8. Why do you not use machine translation in your translation work? (multiple response, n = 39)

Our respondents provide post-editing in 44 unique language directions, the most prominent, again, being English into German (14) and German into English (9), followed by English into Greek and English into Croatian (5 each). English is the dominant source language, accounting for 52 out of the total 99 post-editing language directions indicated. This will have implications for target-language content produced with the help of MT, e.g., when translating into gender-marking languages like German, or when the age of the content author is important – some commercial MT systems have been shown to “make the translated text seem produced by subjects more male and older than what they are” (Bianchi et al., 2023).

No respondent reported having decreased their use of MT recently. 55% saw no change in their MT use, while 45% noticed an increase, justified by a variety of reasons, including:

- recent improvements in MT quality (17 respondents),
- increased workload and staff reduction (17),
- recent access to secure MT (3).

The integration and use of MT within the translation cycle is not without its critics. The question is not whether MT brings any advantages, as these have been widely identified and catalogued. The issue is frequently linked to a constant expectation of higher productivity and lower prices against a decreased inclusion of input from translators in the discourses shaping MT development and integration (Ragni and Vieira, 2022). Instead of the augmented translator supported in her task by tools (Lommel, 2018), as was the case with CAT, many translators experience MT as a disruptive force, endangering their status and agency (Moorkens, 2020). Moreover, mostly or uniquely doing post-editing is linked with concerns about deskilling professionals, as well as decreased motivation and job satisfaction, an aspect which was mentioned by our respondents, and of which academia is also aware (Rothwell et al., 2023). These tensions can be at the root of the discrepancy we noticed in our data analysis.

To further complicate things, new developments in technology and working practices make it increasingly difficult to clearly distinguish between editing TM matches and post-editing MT output (Sánchez-Gijón et al., 2019), the latter being also occasionally presented in the form of a TM.

There also appears to be controversy over the term *post-editing* as expressed in the answers provided by some respondents. In the questionnaire, we provided the fairly broad definition from the ISO 17100:2015 standard, according to which post-editing means to “edit and correct machine translation output” (ISO, 2015, p. 2). However, the standard also includes a note with the definition, according to which using MT suggestions within a CAT tool does not fall under the term post-editing – we did not include this note in the questionnaire, but some participants felt anyway that not all their activities that involve editing and correcting MT output should be classified as post-editing. In the more recent ISO 18587:2017 standard on post-editing (ISO, 2017), the aforementioned note was explicitly removed from the definition of the term, thus broadening the scope of what can be seen as post-editing.

Debate on how workflows involving MT should be called is unlikely to cease, given the increasing relevance of the service and the notions associated with it regarding, among others, productivity pressures, fair pay, translation quality, or work satisfaction. While *translation memory* is a descriptive, motivated, uncontroversial term, *machine translation* assumes (wrongly) that both the process and the product of machines are identical to human ones. When compared to the performance and output of professional translators in optimal workflows, they are not. Yet not only is this controversial term likely to be kept alive by the language services industry, but a similarly misleading sibling *machine interpreting* is gaining ground. This is (also) how professions requiring expertise are ‘deprofessionalised’ in the public perception. The same is true for the term *post-editing*, which implies that the activity consists in making (perhaps even monolingual) edits to an otherwise finished translation product instead of using MT as a tool in a complex and cognitively demanding translation process.

4.2 Use of speech technologies

The use of speech technologies, specifically ASR and TTS, is far less common among our respondents compared to MT (see Table 9). This is not surprising and in line with general industry use (see ELIA et al., 2023).

In our survey, 2 respondents (2%) report currently using ASR. Both use ASR occasionally, and one links use to low MT quality for one of their working language directions (Arabic into Spanish). This association is also made by a third respondent who used ASR in the past but feels that current high-quality MT output renders ASR inefficient. Therefore, we cannot confirm the high uptake identified by Liyanapathirana et al. (2019). Only one person reported still using dictation to a typist or voice recording device, for drafting purposes.

Dictation to typist/recording device	Drafting	1%	
Automatic speech recognition	Translation	2%	
Reading aloud	Translation	17%	25% of respondents
	Post-editing	6%	
	Revision	16%	
	<i>Other</i>	1%	
Text-to-speech synthesis	Translation	3%	6% of respondents
	Revision	3%	
	<i>Other</i>	1%	

Table 9. Summary of voice-enhanced working practices (multiple response items)

One in four respondents (23) said they read aloud, to themselves or others, the texts they are working on and this activity is performed either daily or regularly by the majority. 17% read aloud for translation purposes, 16% for revision and 6% for post-editing. Of all respondents, only 6% use TTS to have their computer read aloud the text for them, albeit mostly occasionally, and out of these, only one also reported reading texts aloud themselves. It could be that those who are reading texts out loud are not aware of ways to integrate TTS into their workflows or, more likely, that their current tech set-up does not support seamless TTS. When

using TTS, the motivation is linked to a perceived ease in revising, and a way to overcome a certain bias in written self-revision. Moreover, this variation in mode of input during revision is also linked to identifying issues linked to fluency, typos, or inconsistencies. Specifically, the reasons given by the respondents include:

- “To vary how I look at texts when revising them”;
- “Checking the quality of alternative texts (easy to grasp when listening?)”;
- “I find it easier to catch mistakes when a text is read aloud and it also helps me not overwork sentences”; and
- The computer “reads what I actually wrote, not what I am convinced I wrote”.

While the above motivations for using speech-enhanced scenarios during translation are solid, further uptake will depend on a combination of technology maturity, seamless integration into existing tools, and individual flexibility. Preference for a certain mode of working may be as important as availability and awareness of speech tools, but increasing pressure to use MT may motivate more translators to look for additional technologies that help deal with MT-specific challenges.

On the other hand, availability of said technologies could also play an important role – especially in an institutional context, where the introduction of new tools can be subject to high IT security requirements – despite a “clear and decisive inclination towards the uptake of the latest technology in institutional practice” (Svoboda and Sosoni, 2022, p. 85).

4.3 Priorities regarding technology and training

One item in our questionnaire asked respondents to specify any priorities they might have in terms of implementing new technologies or receiving technology-related training. Our findings mirror those by Svoboda and Sosoni (2022) in that roughly half (47% in our case) of those responding to this item ($n = 34$) specifically mention MT or post-editing. This makes MT / post-editing the most frequent topic by far and points to an awareness of their growing importance, paired with existing training gaps. Such gaps will continue to widen as new communication needs are identified and included in translators’ purview, but for which the current state of MT is insufficient or even counter-productive, as some respondents highlighted in the general survey comments section: “Machine translation output currently works against the aim to write in plain language and simplify complex sentences.”

Our results therefore show that MT training could be emphasised more, especially when combined with the fact that over half of the total respondents indicated not having received any formal training on MT or post-editing. The notion that institutions introducing MT will frequently offer related training to their employees as identified by Svoboda and Sosoni (2022) might hold true for big international organisations like the EU institutions, but when we look at our entire dataset, 74% of respondents state that MT is integrated into workflows at their institution, while only 30% report having received formal training on MT / post-editing provided by their employer. There is a highly significant difference here between international and national institutions, with international employers being far more likely to provide training for staff ($\chi^2 = 11.849$, $df = 1$, $p = 0.0006$).

As regards further training and integration priorities, interestingly, 7 respondents (21%) mention speech technologies, 6 of whom are not currently using either ASR or TTS; one person is already using TTS but would like to integrate ASR into their work.

Even if a minority, 3 respondents expressed a strong desire for training to return to human-centric approaches, reasoning that the current focus on technology can be demoralising and lead to loss of essential linguistic skills. As one respondent stated,

Article I. “We need to focus on translators for a change, not the tools - they are fine and very helpful but we need to look at the people actually using them, how they are affected and motivated to do their job. If it goes the way it goes now - there will not be many linguists interested in this job, therefore the quality of translation will suffer which is the case already.”

The fact that well over half of our respondents (63%) chose not to answer the question on technology training and implementation priorities can be seen as supporting this view of technology.

5 Limitations

Firstly, our survey cannot claim representativeness for institutional translation at large, since we obtained responses from participants self-selected through snowball sampling rather than a random sample. Our sampling method also explains the strong representation of the European Commission as well as German national institutions among respondents. However, since our dataset is otherwise quite diverse, we believe that the findings are still valuable for shedding light on technological aspects that are important to institutional translators beyond individual organisations.

Secondly, the survey structure could have been optimised to address the fact that some of the respondents use MT to aid them in their day-to-day translation work, but do not consider that to be post-editing. On the other hand, this discrepancy became an interesting part of our data analysis.

6 Conclusions

The growing range of translation and content creation tasks which need to be completed by institutional translators is a constant challenge that some of the latest technologies are expected to help with. Our data shows that, to some extent, this is the case, as MT is increasingly used to speed up processes and help with tasks other than translation. Nevertheless, MT output still requires expert post-editing, and improving MT quality through domain-specific adaptation is just a basic, though still essential step to make the technology useful to institutional translators rather than replace them. For some tasks, such as clear writing and content simplification, MT remains an obstacle, rather than a help. Continued MT training is still necessary, especially in smaller, national institution teams, where new technology roll-outs do not seem to benefit from the same customisation and support as in the larger international organisations. As detailed above, we identified a strong relationship between having access to a customised MT system and a positive opinion about MT quality. This seems particularly relevant given that 17 respondents report not using MT because of its unsatisfactory quality. If we combine these findings with the fact that 16% of our respondents did not know if the MT solution they were using in-house was customised or not, and that those who have received MT training are less likely to use free online MT, we see the need for further training and raising of awareness.

Speech technologies, on the other hand, are used much less, although there is clear interest in training and implementation. Given the still unpredictable nature of MT, they have potential for sustaining quality and productivity of translators, but they also need to be seamlessly integrated into current work environments to prevent them from becoming just another technology which detracts from the translation task instead of assisting it. Additionally, translators need training, so that they are aware of the solutions available to them. 25% of our respondents read aloud when translating, revising, or post-editing but only 6% use TTS to automate this process, which indicates a possible lack of awareness of the technical solutions available for an already familiar task.

Overall, the attitude towards technology that can be identified in the answers to the open-ended questions is the respondents' ultimate objective: to produce an appropriate translation using all the available tools. Talk of how technologies 'simplify' the complex expert task of translating should be kept to a minimum, particularly by non-translators, and technology adoption to cope with increasing workloads should not be confused with technology adoption because of its intrinsic quality and value. The prospect of deskilling professional translators by reducing their role and autonomy within ever-more-complex workflows is an urgent issue to address.

Acknowledgements

We would like to thank the IAMLADP UCG partners who supported the distribution of this survey. Special thanks go to everyone who took time to complete this survey and provide such enriching perspectives on this topic. We would also like to thank our anonymous reviewers for their comments, which helped us improve this article.

References

- Bianchi, Federico, Tommaso Fornaciari, Dirk Hovy, & Debora Nozza (2023). Gender and Age Bias in Commercial Machine Translation. In Helena Moniz & Carla Parra Escartín (Eds.), *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation* (pp. 159–184). Springer International Publishing. https://doi.org/10.1007/978-3-031-14689-3_9
- Bowker, Lynne & Jairo Buitrago Ciro (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited. <https://doi.org/10.1108/9781787567214>
- Brockmann, Justus, Claudia Wiesinger, & Dragoş Ciobanu (2022). Error Annotation in Post-Editing Machine Translation: Investigating the Impact of Text-to-Speech Technology. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 251–259. <https://aclanthology.org/2022.eamt-1.28>
- Cadwell, Patrick, Sheila Castilho, Sharon O'Brien, & Linda Mitchell (2016). Human factors in machine translation and post-editing among institutional translators. *Translation Spaces*, 5(2), 222–243.
- Canfora, Carmen, & Angelika Ottmann (2020). Risks in neural machine translation. *Translation Spaces*, 9(1), 57–77.
- Chereji, Raluca (2024). What makes a medical translator? A survey on medical translators' profiles, work-related challenges and use of Computer-Assisted Translation and Automatic

- Speech Recognition tools. *The Journal of Specialised Translation*, 42, 39–63. <https://doi.org/10.26034/cm.jostrans.2024.5979>
- Ciobanu, Dragoş (2014). Of dragons and speech recognition wizards and apprentices. *Tradumàtica*, 12, 524–538.
- Ciobanu, Dragoş (2016). Automatic speech recognition in the professional translation process. *Translation Spaces*, 5(1), 124–144.
- Ciobanu, Dragoş, Valentina Ragni, & Alina Secară (2019). Speech Synthesis in the Translation Revision Process: Evidence from Error Analysis, Questionnaire, and Eye-Tracking. *Informatics*, 6(4)(51), Article 4. <https://doi.org/10.3390/informatics6040051>
- DePalma, Donald (2014, July). Free machine translation can leak data. *Tcworld Magazine*. <https://www.tcworld.info/e-magazine/translation-and-localization/free-machine-translation-can-leak-data-516/>
- Dragsted, Barbara, Inger M. Mees, & Inge Gorm Hansen (2011). Speaking your translation: Students' first encounter with speech recognition technology. *The International Journal for Translation & Interpreting Research*, 3(1), 10–43.
- ELIA, EMT, EUATC, FIT EUROPE, GALA, LIND, & Women in Localization (2023). *2023 European Language Industry Survey. Trends, expectations and concerns of the European language industry*. <https://elis-survey.org/wp-content/uploads/2023/03/ELIS-2023-report.pdf>
- European Parliament, European Council of Ministers, European Commission, & European Court of Justice (2019). *New Technologies and Artificial Intelligence in the field of language and conference services*. <https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/eu-host-paper-new-technologies-and-artificial-intelligence-field-language-and-conference>
- Farrell, Michael (2023). Do translators use machine translation and if so, how? Results of a survey among professional translators. In Joss Moorkens & Vilelmini Sosoni (Eds.), *Proceedings of the 44th Translating and the Computer (TC44) conference* (pp. 49–60). <https://www.asling.org/tc44/wp-content/uploads/TC44-luxembourg2022.pdf>
- Froeliger, Nicolas, Alexandra Krause, & Leena Salmi (2022). Institutional translation – EMT Competence Framework and beyond. In Tomáš Svoboda, Łucja Biel, & Vilelmini Sosoni (Eds.), *Institutional Translator Training*. Routledge.
- Gaspari, Federico, Hala Almaghout, & Stephen Doherty (2015). A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives*, 23(3), 333–358. <https://doi.org/10.1080/0907676X.2014.979842>
- ISO (2015). *ISO 17100:2015—Translation services – Requirements for translation services*. <https://www.iso.org/standard/59149.html>
- ISO (2017). *ISO 18587:2017—Translation services – Post-editing of machine translation output – Requirements*. <http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/29/62970.html>
- Jayes, Thomas (2023). Conference interpreting and technology: An institutional perspective. In Gloria Corpas Pastor & Bart Defrancq (Eds.), *Interpreting Technologies – Current and*

- Kang, Ji-Hae (2014). Institutions translated: Discourse, identity and power in institutional mediation. *Perspectives*, 22(4), 469–478. <https://doi.org/10.1080/0907676X.2014.948892>
- Kang, Ji-Hae (2019). Institutional translation. In M. Baker & G. Saldanha (Eds.), *Routledge Encyclopedia of Translation Studies* (pp. 256–261). Routledge.
- Koskinen, Kaisa (2008). *Translating institutions: An ethnographic study of EU translation*. St. Jerome.
- Koskinen, Kaisa (2011). Institutional translation. In Yves Gambier & Luc van Doorslaer (Eds.), *Handbook of Translation Studies* (pp. 54–60). John Benjamins.
- Lafeber, Anne (2012). Translation Skills and Knowledge – Preliminary Findings of a Survey of Translators and Revisers Working at Inter-governmental Organizations. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, 57(1), 108–131. <https://doi.org/10.7202/1012744ar>
- Lafeber, Anne (2022). Skills and knowledge required of translators in institutional settings. In Tomáš Svoboda, Łucja Biel, & Vilelmini Sosoni (Eds.), *Institutional Translator Training*. Routledge.
- Liyanapathirana, Jeevanthi, Pierrette Bouillon, & Bartolomé Mesa-Lao (2019). Surveying the potential of using speech technologies for post-editing purposes in the context of international organizations: What do professional translators think? *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, 149–158. <https://aclanthology.org/W19-6728>
- Lommel, Arle (2018). Augmented Translation: A New Approach to Combining Human and Machine Capabilities. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, 5–12. <https://aclanthology.org/W18-1905>
- Macken, Lieve, Daniel Prou, & Arda Tezcan (2020). Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process. *Informatics*, 7(2), Article 2. <https://doi.org/10.3390/informatics7020012>
- Martín Ruano, M. Rosario (2019). Legal and institutional translation. In *The Routledge Handbook of Spanish Translation Studies*. Routledge.
- Mesa-Lao, Bartolomé (2014). Speech-Enabled Computer-Aided Translation: A Satisfaction Survey with Post-Editor Trainees. *Proceedings of the EACL 2014 Workshop on Humans and Computer-Assisted Translation*, 99–103.
- Moorkens, Joss (2020). “A tiny cog in a large machine”: Digital Taylorism in the translation industry. In *Translation Spaces* (Vol. 9, Issue 1, pp. 12–34). John Benjamins. <https://doi.org/10.1075/ts.00019.moo>
- Mossop, Brian (1988). Translating institutions: A missing factor in translation theory. *TTR : Traduction, Terminologie, Rédaction*, 1(2), 65–71. <https://doi.org/10.7202/037019ar>
- Nimdzi Insights. (2023). *Nimdzi Language Technology Atlas*. <https://www.nimdzi.com/language-technology-atlas/>

- Orlando, Marc, Liao, Sixin, & Jan-Louis Kruger (2024). *Translation and Interpreting technologies and their impact on the industry*. Macquarie University.
- Posit team (2023). *RStudio: Integrated Development Environment for R* [Computer software]. Posit Software, PBC. <http://www.posit.co/>
- Pym, Anthony (2008). The use of translation in international organizations. In Harald Kittel, Armin Paul Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert, & Fritz Paul (Eds.), *Übersetzung Translation Traduction: An international Encyclopedia of Translation Studies* (Vol. 1, pp. 85–92). De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110137088.1.2.85>
- R Core Team (2023). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ragni, Valentina, & Lucas Nunes Vieira (2022). What has changed with neural machine translation? A critical review of human factors. *Perspectives*, 30(1), 137–158. <https://doi.org/10.1080/0907676X.2021.1889005>
- Rios Gaona, Miguel Angel, Justus Brockmann, Claudia Wiesinger, Raluca Chereji, Alina Secară, & Dragoş Ciobanu (2024). Bayesian Hierarchical Modelling for Analysing the Effect of Speech Synthesis on Post-Editing Machine Translation. In Xingyi Song, Edward Gow-Smith, Carolina Scarton, Vera Cabarrão, Konstantinos Chatzitheodorou, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Rachel Bawden, Víctor M. Sánchez-Cartagena, Barry Haddow, Diptesh Kanojia, Mary Nurminen, Helena Moniz, Mikel Forcada, Chris Oakley (Eds.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation, Volume 1: Research And Implementations & Case Studies* (pp. 455–68). <https://eamt2024.github.io/proceedings/vol1.pdf>
- Rossi, Caroline, & Jean-Pierre Chevrot (2019). Uses and perceptions of machine translation at the European Commission. *The Journal of Specialised Translation*, 31, 177–200.
- Rothwell, Andrew, Joss Moorkens, María Fernández-Parra, Joanna Drugan, & Frank Austermuehl (2023). *Translation Tools and Technologies* (1st ed.). Routledge. DOI: 10.4324/9781003160793
- Sánchez-Gijón, Pilar, Joss Moorkens, & Andy Way (2019). Post-editing neural machine translation versus translation memory segments. *Machine Translation*, 33, 31–59.
- Svoboda, Tomáš, Łucja Biel, & Krzysztof Łoboda (Eds.). (2017). *Quality aspects in institutional translation*. Language Science Press.
- Svoboda, Tomáš, Łucja Biel, & Vilelmini Sasoni (Eds.). (2022). *Institutional Translator Training* (1st ed.). Routledge. <https://doi.org/10.4324/9781003225249>
- Svoboda, Tomáš, & Vilelmini Sasoni (2022). Institutional translator training in language and translation technologies. In Tomáš Svoboda, Łucja Biel, & Vilelmini Sasoni (Eds.), *Institutional Translator Training*. Routledge. <http://doi.org/10.4324/9781003225249>
- Vardaro, Jennifer, Moritz Schaeffer, & Silvia Hansen-Schirra (2019). Translation Quality and Error Recognition in Professional Neural Machine Translation Post-Editing. *Informatics*, 6(3).
- Wiesinger, Claudia, Justus Brockmann, Alina Secară, & Dragoş Ciobanu (2022). Speech-enabled machine translation post-editing in the context of translator training. In Michał

- Kornacki & Gary Massey (Eds.), *Contextuality in Translation and Interpreting. Selected Papers from the Łódź-ZHAW Duo Colloquium on Translation and Meaning 2020–2021* (Vol. 70). Peter Lang.
- Zapata, Julián, Sheila Castilho, & Joss Moorkens (2017). Translation Dictation vs. Post-editing with Cloud-based Voice Recognition: A Pilot Experiment. *Proceedings of Machine Translation Summit XVI: Commercial MT Users and Translators Track*, 116–129. <https://aclanthology.org/2017.mtsummit-commercial.13>
- Zapata, Julián, Alina Secară, & Dragoş Ciobanu (2023). Past, present and future of speech technologies in translation—Life beyond the keyboard. In Joss Moorkens & Vilemini Sisoni (Eds.), *Proceedings of the 44th Translating and the Computer (TC44) conference* (pp. 16–25). Editions Tradulex.
- Zaretskaya, Anna, Gloria Corpas Pastor, & Miriam Seghiri (2015). *Translators' Requirements for Translation Technologies: A User Survey*. The 7th International Conference of the Iberian Association of Translation and Interpreting Studies (AIETI). New Horizons in Translation and Interpreting Studies, Málaga, Spain.

How can Paidiom improve the neural machine translation of multiword expressions?

Carlos Manuel Hidalgo-Ternero

Universidad de Málaga
IUITLM

Malaga, Spain

cmhidalgo@uma.es

Francisco Javier Lima-Florido

Universidad de Málaga
IUITLM

Malaga, Spain

fco.javier.lima@uma.es

Abstract

In this paper we present research results with Paidiom, a text-preprocessing algorithm designed for 1) converting discontinuous multiword expressions (MWEs) into their continuous forms and 2) translemmatising them, i.e., converting source-text MWEs into their target-text equivalents, to improve the performance of current neural machine translation (NMT) systems. To test its effectiveness, an experiment with VIP (Corpas Pastor, 2021), Google Translate and DeepL NMT systems was carried out in the ES>EN translation direction with Spanish Verb-Noun Idiomatic Constructions (VNICs). The performance of Paidiom was compared both to our previous algorithm (gApp) and to manual conversion (our gold standard). The promising results yielded by this study, the first one analysing Paidiom's performance, shed some light on new avenues for enhancing MWE-aware NMT systems.

1 Introduction

The recent emergence of neural networks in machine translation has represented a real breakthrough, Neural Machine Translation (NMT), which has resulted in a considerable qualitative leap compared to previous ruled-based and statistical models (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Wang et al., 2022).

Despite these advances, NMT systems still have an important Achilles' heel: MWEs. Besides their quintessential problematic features such as syntactic anomaly, non-compositionality, diasystematic variation and ambiguity, a further challenge arises for NMT: MWEs do not always consist of adjacent tokens (e.g., *You need to keep those things in mind.*), which seriously hinders their automatic detection and translation (Constant et al., 2017; Copras Pastor, 2013; Ramisch & Villavicencio, 2018; Rohanian et al., 2019). To overcome the challenges that discontinuous MWEs still pose for even the most robust NMT systems (cf. Colson, 2019; Zaninello & Birch, 2020), we designed an upgraded, freely-available algorithm, called Paidiom¹, which is able not only to automatically convert discontinuous MWEs into their continuous form (analogously to our previous algorithm gApp [see Hidalgo-Ternero, 2021 and 2024; Hidalgo-Ternero and Copras Pastor, 2020, 2024a and 2024b; Hidalgo-Ternero and Zhou-Lian, 2022]) but also to translemmatise them, i.e., to directly convert MWEs into their target-text equivalents to improve NMT (see Section 2.3).

¹ A detailed explanation of Paidiom is available through this link: <https://lexytrad.es/paidiom>

Against this background, the current study analyses the neural machine translation of MWEs after Paidiom automatic conversion. To this end, the performance of VIP (Corpas Pastor, 2021), Google Translate and DeepL NMT systems will be examined against 400 cases: 100 discontinuous forms of MWEs (i.e., the original texts), 100 continuous forms after gApp conversion, 100 continuous and translemmatised forms after Paidiom conversion, and 100 continuous and translemmatised forms after manual conversion; the latter constituting our gold standard. The MWEs under study include the following: *haber gato encerrado*, *ser cuatro gatos*, *dormir la mona* and *ganar/costar/pagar cuatro perras*, which will be analysed in Section 5.

The remainder of the paper is structured as follows. Section 2 introduces the text preprocessing algorithm Paidiom, Section 3 examines the MWEs under study, and Section 4 illustrates the research methodology. In Section 5, the algorithm’s precision and recall will first be tested, to assess to what extent Paidiom can enhance the performance of VIP, Google Translate and DeepL’s NMT systems under the challenge of MWE discontinuity in the Spanish-into-English translation direction. A discussion of the results will follow (Section 6). Section 7 provides concluding remarks on how to further test Paidiom following this study’s findings.

2 Overview of Paidiom

The programming language employed for Paidiom was Python 3.7, plus the Spacy library, specialised in performing a wide array of advanced NLP tasks, including non-destructive tokenization, POS tagging, dependency parsing, lemmatisation, and rule-based matching (Honnibal and Montani, 2017). More specifically, the pretrained statistical model for Spanish `es_core_news_sm` was used, i.e., a multi-task convolutional neural network (CNN) trained on WikiNER (Nothman et al., 2017) and UD Spanish AnCora (Martínez Alonso and Zeman, 2016).

Against this background, Paidiom was designed to carry out three main tasks in MWE preprocessing: first the detection of discontinuous MWEs (Section 3.1), secondly the conversion of these MWEs into their continuous forms (these first two tasks are also carried out by gApp) (Section 3.2), and finally the translemmatisation of MWEs to enhance NMT performance (Section 3.3).

2.2 Automatic detection of discontinuous MWEs

The algorithm Paidiom performs a token-based MWE identification. Thus, following a Lexicon Lookup Method (Ramisch and Villavicencio, 2018), it refers to a predefined lexicon of semi-fixed expressions, i.e., those which can undergo any kind of internal morphosyntactic alteration from their canonical form (Sag, Baldwin et al., 2002). These semi-fixed expressions can occur in texts in a discontinuous form, i.e., other elements can appear embedded within the constituents of the MWEs.

As a prior step to the implementation of pattern detection, it was necessary to establish what kinds of unigrams, bigrams or trigrams can occur within (and hence split) the discontinuous form of the MWEs. Thus, we queried two giga-token web-crawled Spanish corpora (`esTenTen18` and `Timestamped JSI web corpus 2014-2021 Spanish`), both accessible through Sketch Engine. The `esTenTen18` corpus comprises over 17 billion words of general (both European and American) Spanish, with a heterogeneous sample in terms of text sources, types and diasystematic varieties (including User Generated Content [UGC]). On the other hand, the

Timestamped JSI web corpus 2014-2021 Spanish contains over 16.4 billion words of news articles obtained from their RSS feeds (Kilgarriff et al., 2004).

Following Hidalgo-Tertero's (2020) corpus-based research methodology, Sketch Engine's Corpus Query Language (CQL) schemas were employed to retrieve both the discontinuous forms of the MWEs under study (henceforward designated as relevant results) as well as other concordances with similar patterns but unrelated to the MWEs (irrelevant results). This will ultimately delimit the necessary restrictions for the Paidiom detection system to optimise its precision and recall. In this sense, several rule-based matching patterns were set within the lexicon so that Paidiom can detect as many relevant results while filtering out as many irrelevant results as possible. These patterns comprise a list of dictionaries, each of which includes the necessary description of both the exact MWE tokens and the constituents that may appear within (and hence split) the MWE, necessary to proceed with the conversion stage.

2.3 Automatic conversion of discontinuous MWEs into their continuous form

Once the detection is completed, the system can proceed with the automatic conversion of discontinuous MWEs into their continuous form. In this way, with a for-loop a first condition is determined if the algorithm matches any of the predefined patterns within the lexicon of MWEs. In this scenario, the system is set to detect the first dictionary of the match (called *pos_ini*, i.e., 'initial position') and the last one (called *pos_fin*, i.e., 'final position') with the gap as those optional elements within that sequence. In this way, the first token within the gap (called *gap1*) would be *pos_ini+1*, and the final one (called *gap3*) would be *pos_fin-2*.

After the delimitation of the gap, the algorithm is set to automatically generate from the beginning of the text up to *pos_ini*, then *pos_fin*, subsequently from *gap1* up to *gap3*, and finally from *pos_fin* up to the end of the document, which results in the whole original text with the MWE in its continuous form as the output. If the first condition is not met, i.e., if none of the predefined patterns within the lexicon of MWEs is matched, no MWE is converted to its continuous form. This process takes place iteratively until all discontinuous MWEs in the text are converted.

2.4 Automatic translemmatisation of MWEs

Before delving into the notion of MWE *translemmatisation* with Paidiom, we first need to explain the underlying concept of *translemma*. A *translemma* is a "bitextual unit of any type or level consisting of the same content and two formal manifestations that are different but mutually solidary and whose existence depends on the overall relationship of equivalence underlying each source-text/target-text binomial"² (Santoyo and Rabadán 1991: 322).

In this context, we will now describe as *translemmatisation* the process of converting a source-text lexical unit into its target-text equivalent, both of which conform to a *translemma* or bitextual unit. Against such a background, our main hypothesis is that NMT performance can

² "Unidad bitextual de cualquier tipo o nivel constituida por un mismo contenido y dos manifestaciones formales diferenciadas pero solidarias y cuya existencia depende de la relación global de equivalencia subyacente a cada binomio textual TO-TM." [original version in Spanish, the translation into English is ours]

be considerably improved if MWEs are not only converted into their continuous form but also translemmatised prior to NMT.

Within Paidiom, the translemmatisation process starts with a mapped data structure where the MWEs studied in the source language (and now in their continuous form) are linked to their corresponding equivalent MWE in the target language. This data structure serves as a database in a dictionary format. As the ES/EN *translemmas* were already set, the next step was to update the *pos_ini* and *pos_fin* from the discontinuous-to-continuous conversion to each corresponding MWE in its continuous form. After this, the algorithm copies the text from the beginning to the *pos_ini* of the first MWE, creating a new converted text. Later, it makes a query in the database to obtain the corresponding MWE in the target language and concatenates the copied text with the target MWE. This is produced in a loop to convert each MWE starting from the *pos_fin* of the last MWE occurrence. At the end of the loop, the remaining part of the text not containing any additional MWE to be translemmatised is linked to the previously converted text, resulting in a final output in which all MWEs in the lexicon have been converted into their continuous form and translemmatised to improve NMT, as shown in Section 6 (*Results*).

3 The MWEs under study

To test the effectiveness of Paidiom in the ES>EN translation direction, for the present study we selected the following four Spanish MWEs:

Haber gato encerrado: ‘Haber causa o razón oculta o secreta, o manejos ocultos.’ (‘there is sth secret or hidden.’) (DLE, 2023); ‘Haber algo oculto.’ (‘there is sth hidden.’) (DFDEA, 2017: 365).

Primary correspondence(s) in English: *there is a catch, there is something fishy going on*

(1) Aquí hay gato encerrado.

EN: lit. ‘There is a cat locked in here.’

‘There’s something fishy going on here.’

Ser cuatro gatos: ‘Poca gente y sin importancia.’ (‘few unimportant people.’) (DLE, 2022); ‘Muy poca gente.’ (‘very few people.’) (DFDEA, 2017: 364).

Primary correspondence(s) in English: *there is (just) a small bunch of people*

(2) No eran más que cuatro gatos en la manifestación.

EN: lit. ‘There weren’t more than four cats at the demonstration.’

‘There was only a small bunch of people at the demonstration.’

Dormir la mona: ‘Dormir después de una borrachera o del consumo de drogas.’ (‘to sleep after getting drunk or after taking drugs.’) (DFDEA, 2017: 530).

Primary correspondence(s) in English: *to sleep sth off*

(3) Después de emborracharse, se fueron a dormir la mona en una esquina.

EN: lit. ‘After getting drunk, they went to sleep the female monkey at a corner.’

‘After getting drunk, they went to sleep off the hangover at a corner.’

Ganar/costar/pagar... cuatro perras: ‘[ganar/costar/pagar...] una cantidad insignificante de dinero’ (‘[earn/cost/pay...] an insignificant amount of money.’) (DFDEA, 2017: 646)

Primary correspondence(s) in English: *[to earn, to cost, to pay] peanuts.*

(4) En este bar los camareros ganan cuatro perras.

EN: lit. ‘In this bar waiters earn four bitches.’

‘In this bar waiters are paid peanuts.’

Following Ramisch’s (2015) taxonomy, the MWEs *haber gato encerrado* and *dormir la mona* are classed as *idiomatic expressions*, since they have a non-compositional meaning (which is why they are also defined as *semantically non-decomposable idioms* or *SNDIs* [Bargman and Sailer, 2018]). The MWEs *ser cuatro gatos* and *ganar/costar/pagar... cuatro perras* are collocations consisting of a verb, with a literal meaning, and a noun-phrase idiom (*cuatro gatos* and *cuatro perras*, respectively). In this context, we deemed it necessary to include the verbs with which these idioms most commonly collocated (see Figures 1 and 2 for the patterns “Verb + *cuatro gatos*” and “Verb + *cuatro perras*” in esTenTen18), since, with other verbs and verb phrases such as *alimentar* (‘feed’), *vivir con* (‘live with’), *cuidar (de)* (‘take care of’), *hacerse cargo de* (‘be in charge of’) etc., both *cuatro gatos* and *cuatro perras* tend to appear in the corpus with their literal meanings (‘four cats’ and ‘four bitches’, respectively), hence the need to examine the collocation as a whole.

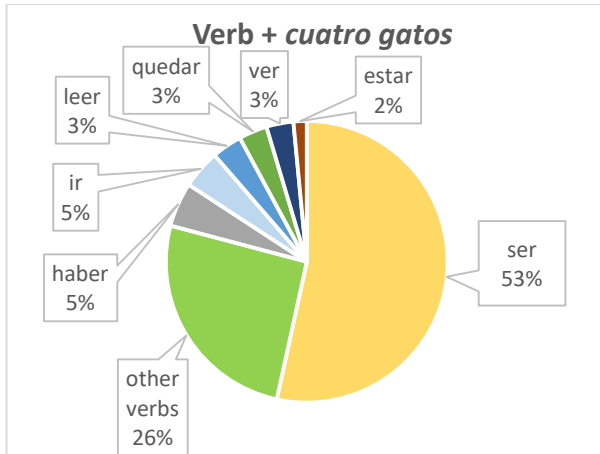


Figure 5. Verbs collocating with *cuatro gatos*

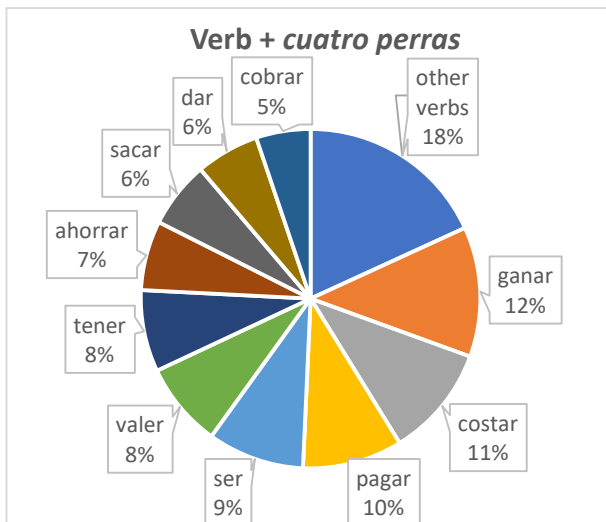


Figure 6. Verbs collocating with *cuatro perras*

With regard to their morphosyntactic structure, the four MWEs belong to the category of *verb-noun idiomatic constructions* (VNICs) (Fazly et al., 2009) as they all consist of a verb and a noun in its direct object position. Concerning their fixedness, following the taxonomy of Parra Escartín et al. (2018) for Spanish MWEs, they can be classified as *flexible*, since other elements can appear embedded within the constituents of the MWEs. Finally, considering the nature of their constituents, they are *zoo-logisms*, i.e., MWEs containing terms that refer to animal names (called *zoonyms*). In this regard, we decided to analyse specifically (partly) idiomatic expressions because their non-compositional meaning makes them potentially easier to detect and translate by NMT systems when all the constituents of the MWE are contiguous, as shown in Table 1 (Section 1).

As regards the frequency of the different MWEs in the corpus esTenTen18, Table 2 summarises their raw frequency (Column 3), their normalised frequency per million tokens (Column 4) and the percentage (Column 5) appearances of these MWE in their continuous (C.) or discontinuous (D.) forms in relation to their total occurrences (T.).

MWE	Form	R. F.	N. F.	%
Haber gato encerrado	C.	2299	0.12	89.6%
	D.	267	0.01	10.4%
	T.	2566	0.13	
Ser cuatro gatos	C.	1290	0.07	74.8%
	D.	434	0.02	25.2%
	T.	1724	0.09	
Dormir la mona	C.	779	0.04	96.2%
	D.	31	>0.01	3.8%
	T.	810	0.04	
Ganar/costar/pagar... cuatro perras	C.	284	0.01	74.3%
	D.	98	0.01	25.7%
	T.	382	0.02	

Table 10. appearances of the MWEs in esTenTen18

As shown in Table 2, these four MWEs mainly appear throughout the corpus in their continuous forms (89.6% for *haber gato encerrado*, 74.8% for *ser cuatro gatos*, 96.2% for *dormir la mona*, and 74.3% for *ganar/costar/pagar... cuatro perras*). Continuous occurrences are 3 times more frequent than discontinuous ones. In this context, in this study we intend to test our main hypothesis: that Paidiom can improve NMT performance by converting MWEs into their canonical state (i.e., their continuous form) and by translemmatising them.

4 Methodology

This section presents the research methodology employed to assess to what extent Paidiom can improve the performance of VIP, Google Translate and DeepL NMT systems in the ES>EN translation direction. Analogously to Hidalgo-Terner (2020), the concordances containing the discontinuous MWEs under study were also retrieved from the Spanish corpora esTenTen18 and Timestamped JSI web corpus 2014-2021 Spanish. To analyse the different English translations offered by VIP, Google Translate and DeepL for the source-text MWEs, we used the Sketch Engine corpora enTenTen20 (38.1 billion words) and Timestamped JSI web corpus 2014-2021 English (60.4 billion).

Despite the challenges still posed by ubiquitous source-text error, noise and out-of-vocabulary tokens in user-generated content (UGC) for even the most robust NMT systems (Belinkov and Bisk, 2018; Lohar et al., 2019), a heterogeneous sample in terms of language varieties, text sources and types (including UGC) was selected for the analysis to alleviate sampling bias, which could otherwise originate from exclusively examining canonical NMT training data for these MWEs. In this way, a total of 400 cases were analysed, comprising 100 discontinuous forms (i.e., the texts in their original version), 100 continuous forms (i.e., after gApp conversion), 100 continuous and translemmatised forms after Paidiom conversion and

100 continuous and translemmatised forms after the manual conversion (i.e., our gold standard) of the MWEs *haber gato encerrado*, *ser cuatro gatos*, *dormir la mona*, and *ganar/costar/pagar... cuatro perras*. Besides these relevant results, for each MWE 25 irrelevant results were compiled, to calculate, in a first stage, both the precision and recall of this system, considering all the constituents of the MWE.

Once both parameters were quantified, in a second stage, the performance of VIP, Google Translate and DeepL for the different concordances were classified within the four main categories: before conversion, after automatic conversion with gApp, after automatic conversion and translemmatisation with Paidiom, and after manual conversion and translemmatisation (i.e., the gold standard). The NMT outputs for these different scenarios were then manually assessed with an MT evaluation method based on directly expressed judgements (*DEJ-based evaluation method*, cf. Chatzikoumi, 2020). We decided to use human evaluation for our study given the obstacles that automatic metrics present for specifically evaluating the phenomenon of idiom translation:

Global metrics, such as BLEU (Papineni et al., 2002), consider the full translation, and thus, the effects of idiom translation are overshadowed. Previous efforts on targeted evaluation isolated the idiom translation using word alignments (Fadaee et al., 2018) or word edit distance (Zaninello and Birch, 2020). These approaches measure the accuracy of the idiom translation but do not account for literal translation errors. Shao et al. (2018) proposed a method for estimating the frequency of such errors, but this requires the creation of language-specific handcrafted lists (i.e., blocklists) with words that correspond to literal translation errors. (Baziotis et al., 2022: 1)

For this reason, in our study three professional ES/EN translators, with between 4 and 10 years of experience, were selected as annotators to directly express judgement on the translation quality of the different MT outputs using a binary scale: 1 (good) or 0 (bad). After they submitted their judgements, final decisions on the acceptability (or not) of each specific target text were made on a majority basis: for instance, if 2 or 3 of the translators had judged an MT output as *good*, then this output was also finally categorised as *good* for our study, and vice versa. When judging MT quality, they were specifically instructed to focus exclusively on the phenomenon of MWE discontinuity for the MWEs under study, i.e., whether the ST idiom was accurately conveyed in the TT and, where applicable, whether the element causing the discontinuity was still appropriately rendered in the TT. As the ST was altered by gApp (joining the discontinuous MWE) and by Paidiom (partially translemmatising the ST idiom), they were also instructed to consider whether this alteration in the structure of the input sentence caused any additional error in the rest of the text that was not already present in the TT of the original (discontinuous) scenario, if so, they had to consider the MT output as *bad* (since the ST alteration caused an unprecedented error in the TT). If not, morphological, syntactic, and/or orthotypographic divergences or source-text/translation inaccuracies affecting other elements in the target text were not rated by the human annotators if they were completely unrelated to the phenomenon of MWE discontinuity for the MWEs under study.

5 Results

In this section, the results will be examined and shown at two stages: first, Paidiom precision and recall for each of the MWEs under study will be presented to evaluate to what extent this system can enhance the performance of VIP, Google Translate and DeepL under the challenge of MWE discontinuity in the ES>EN directionality.

As regards Paidiom precision and recall, in the case of *haber gato encerrado* (henceforth presented in the different tables and figures as MWE1), the system automatically converted 24 forms, all of which were true positives. Thus, Paidiom’s precision was 100% (24/24 cases) and its recall 96% (24/25 cases). For the MWE *ser cuatro gatos* (MWE2), Paidiom made 26 conversions, 25 of which were true positives and 1 was a false positive. Therefore, its precision amounted to 96.2% (25/26 cases) and its recall to 100% (25/25). For the idiom *dormir la mona* (MWE3), Paidiom converted all true positives (and no false positive), thus with a precision and recall of 100%. Finally, in the case of *ganar/costar/pagar... cuatro perras* (MWE4), Paidiom performed 27 conversions, 25 of which were true positives and 2 were false positives; thus with a precision of 92.6% (25/27) and recall of 100% (25/25). The precision and recall of Paidiom is presented in Table 3.

	Precision	Recall	F1
MWE1	100%	96%	98%
MWE2	96.2%	100%	98.1%
MWE3	100%	100%	100%
MWE4	92.6%	100%	96.2%
Average	97.2%	99%	98.1%

Table 11. Paidiom precision and recall

After calculating precision and recall, the performance of the different NMT systems was analysed. The improvements in VIP, before and after Paidiom, shown in Figure 3, reveals analogous results for the MWEs under study.

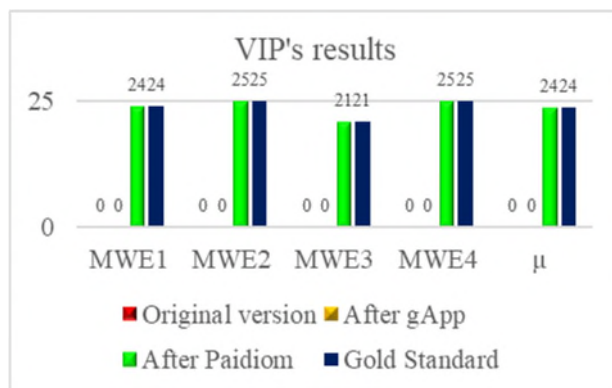


Figure 7. VIP’s results

Figure 3, shows that VIP was unable to detect and translate the different MWEs either in their discontinuous form or in their continuous form after gApp, and that it was only after transliteration that it could offer an average (μ) performance of 96% after Paidiom and the gold standard.

Slightly better results in the discontinuous and continuous forms (after gApp) were obtained with Google Translate (Figure 4), with an average performance of 12% in both the discontinuous and the continuous scenario. Once again, a considerable improvement can be

observed after the translemmatisation of these MWEs both with Paidiom (92% accuracy) and with the gold standard (93.3%).

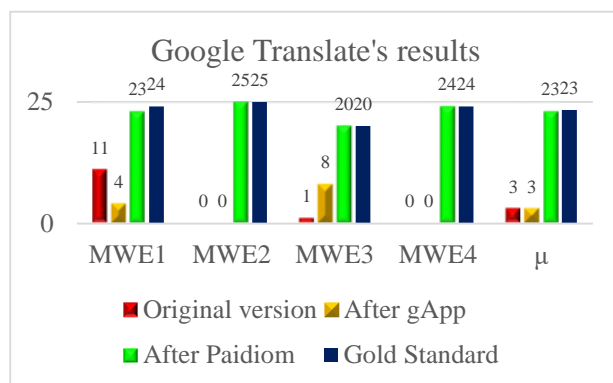


Figure 8: Google Translate's results

Finally, regarding DeepL's accuracy (Figure 5), we can observe how this was the only NMT for which gApp could, on average, improve the performance of the original (i.e., discontinuous) scenario, by 14%. DeepL's accuracy after the translemmatisation was analogous to Google Translate, i.e., 91.2% with Paidiom and 92% with the gold standard.

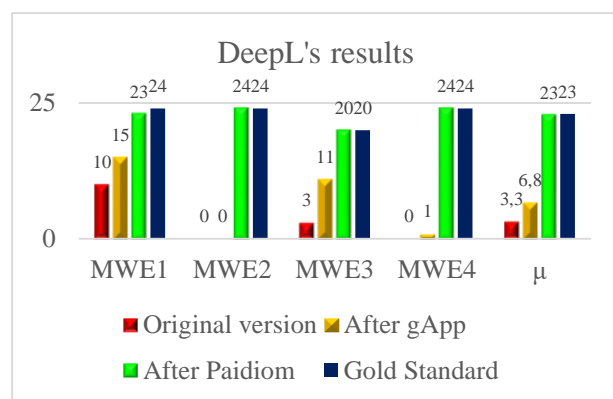


Figure 9: DeepL's results

To illustrate how Paidiom could improve NMT performance, let us observe the following instances with the source-text MWE *haber gato encerrado* translated by Google Translate in its original (i.e., discontinuous) version, after gApp, and after Paidiom (Tables 4 and 5) (the whole sequence is in bold and the MWE is underlined, for illustration purposes).

	KWIC extracts
ST [ES] Original version	[...] circunscribir solo a Cortés lo que López Obrador ha sacado al baile, sería miope. ¿ <u>Hay aquí gato encerrado</u> ? ¿Es ese un anuncio que en el fondo pretende nada más que la favorabilidad que trasnocha a los políticos? [...]

ST [ES] After gApp	[...] circunscribir solo a Cortés lo que López Obrador ha sacado al baile, sería miope. ¿ Hay gato encerrado aquí? ¿Es ese un anuncio que en el fondo pretende nada más que la favorabilidad que trasnocha a los políticos? [...]
ST [ES] After Paidiom	[...] circunscribir solo a Cortés lo que López Obrador ha sacado al baile, sería miope. ¿ Hay something fishy going on aquí? ¿Es ese un anuncio que en el fondo pretende nada más que la favorabilidad que trasnocha a los políticos? [...]

Table 12: Source-text KWIC extracts with *haber gato encerrado* before and after gApp and Paidiom

	KWIC extracts
TT [EN] Original version	[...] to limit only to Cortés what López Obrador has brought to the dance would be myopic. Is there a cat locked up here? Is that an announcement that basically seeks nothing more than the favorability that stays up late for politicians? [...]
TT [EN] After gApp	[...] to limit only to Cortés what López Obrador has brought to the dance would be myopic. Is there a cat locked up here? Is that an announcement that basically seeks nothing more than the favorability that stays up late for politicians? [...]
TT [EN] After Paidiom	[...] to limit only to Cortés what López Obrador has brought to the dance would be myopic. Is there something fishy going on here? Is that an announcement that basically seeks nothing more than the favorability that stays up late for politicians? [...]

Table 13: Google Translate outcomes before and after the conversion of the ST idiom *haber gato encerrado* with gApp and Paidiom

In Table 5 we can observe distinctly different results before and after the automatic conversion of the source-text MWE with Paidiom. Both in the original (i.e., discontinuous) scenario and after gApp, the sequences *¿Hay aquí gato encerrado?* and *¿Hay gato encerrado aquí?* were translated as *Is there a cat locked up here?* Nevertheless, the context (an announcement made by Mexico’s current Prime Minister Andrés Manuel López Obrador) and co-occurrences such as *miope* (‘myopic’) and *en el fondo* (‘deep down’) make us understand that the writer is not referring to any feline³ whatsoever, but he is denouncing that there might be something hidden behind López Obrador’s words. Against this background, it is only after the translemmatisation with Paidiom that Google Translate can deliver an appropriate translation: *Is there something fishy going on here?*

³ Furthermore, the sequence *haber Ø gato encerrado* exclusively has the figurative meaning described in Section 3 (‘there is something fishy going on’). For this sequence to acquire a literal reading, a modifier is needed for *gato encerrado*, for instance, *hay un gato encerrado* (‘there is a cat locked up’) or *hay algún gato encerrado* (‘there is some cat locked up’), among others.

6 Analysis of the results

Global results (Figure 6) show how Paidiom automatic translemmatisation managed to achieve an analogous performance to the gold standard. This is chiefly due to Paidiom’s refined detection system both in terms of final average precision (97.2%) and recall (99%), resulting in an F1 score of 98.1%.

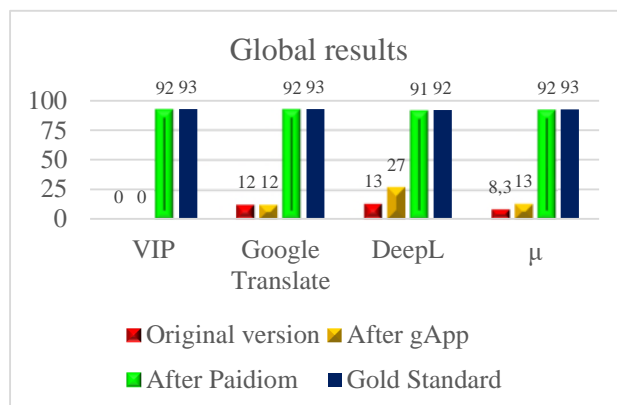


Figure 10: Global results

Global results (Figure 6) also reflect our previous experiments with gApp, in which we proved that NMT of discontinuous MWEs can be improved overall by converting them into their continuous form: in this experiment, NMTs achieved an 8.3% accuracy in the discontinuous form vs. 13% with the gApp continuous form, i.e., an enhancement of 4.7%. Furthermore, global results confirm our initial hypothesis: NMT performance can be considerably improved if MWEs are not only converted into their continuous form but also translemmatised prior to NMT. As shown in Figure 6, the conversion into the continuous form and the translemmatisation with Paidiom led NMT systems to achieve an overall 91.7% accuracy, thus an analogous performance to the gold standard (92.7%). When comparing Paidiom to the original (discontinuous) version, it could, on average, improve NMT by 83.4% (84.4% with the gold standard).

Some of the remaining 8.3% of the global results that could still not be properly detected and translated by the NMT systems, even after the conversion into the continuous form and translemmatisation with Paidiom, are illustrated in Tables 6 and 7.

	KWIC extracts
ST [ES] Original version	¿Para qué zurcir un calcetín? compra unos nuevos ¿hacer bechamel? pero si la venden ya hecha ¿arreglar unas cortinas? si valen como mucho cuatro perras en el Ikea... [...]
ST [ES] After gApp	¿Para qué zurcir un calcetín? compra unos nuevos ¿hacer bechamel? pero si la venden ya hecha ¿arreglar unas cortinas? si valen cuatro perras como mucho en el Ikea... [...]

ST [ES] After Paidiom	¿Para qué zurcir un calcetín? compra unos nuevos ¿hacer bechamel? pero si la venden ya hecha ¿arreglar unas cortinas? si valen peanuts como mucho en el Ikea... [...]
--	--

Table 14. Source-text KWIC extracts with *valer cuatro perras* before and after gApp and Paidiom

	KWIC extracts
TT [EN] Original version	Why mend a sock? buy some new ones make béchamel? but if they sell it already made, fix some curtains? if four bitches are worth at most in Ikea...
TT [EN] After gApp	Why mend a sock? buy some new ones make béchamel? but if they sell it already made, fix some curtains? if they are worth four bitches at most in Ikea...
TT [EN] After Paidiom	Why mend a sock? buy some new ones make béchamel? but if they sell it already made, fix some curtains? Peanuts are worth a lot at Ikea...

Table 15. Google Translate outcomes before and after the conversion of the ST idiom *valer cuatro perras* with gApp and Paidiom

The instances in Table 7 show distinctly different Google Translate outputs before and after the automatic conversion of the source-text MWE with gApp and with Paidiom. In the discontinuous scenario, the sequence *vale como mucho cuatro perras* was translated into *four bitches are worth at most*, where we can observe a literal (and hence inadequate) translation of *cuatro perras* as *four bitches* (with the meaning of ‘four female dogs’). *Cuatro perras* was also incorrectly parsed as the subject of *valer* (‘to cost’) instead of its actual function as a quantitative adverbial modifier (in this sentence, the actual subject of *cost* is *the curtains*, which was omitted in the ST in Spanish). After gApp, *vale cuatro perras como mucho* was translated into *they are worth four bitches at most*, which means that *cuatro perras* was now correctly parsed as the modifier of *valer* but it was still interpreted in its literal (and hence inappropriate) meaning. It was only after Paidiom that *cuatro perras* was correctly translated as *peanuts* (in the sense of ‘small amount of money’). However, it was once again wrongfully parsed as the subject of *costar*, leading to a target text with a completely different meaning from the ST: *peanuts are worth a lot at Ikea*. These results emphasise the fact that, besides the challenges posed by MWE discontinuity and crosslinguistic anisomorphism, there are other obstacles (such as incorrect dependency parsing) that still need to be fully addressed to pave the road towards 100% accuracy in NMT.

7 Conclusion

The findings of our study confirm our initial hypothesis: the Paidiom system can improve NMT performance by converting discontinuous MWEs into their continuous form and by translemmatising them. More specifically, Paidiom is shown to enhance NMT for the MWEs analysed with a final average improvement of 83.4%, which is only 1% lower than our gold standard, i.e., the manual conversion (84.4%).

In this context, the promising results of this study invite us to continue evaluating Paidiom in further experiments to determine to what extent it can also improve NMT performance for other idioms, as well as for other MWE typologies (collocations, verb-particle constructions, etc.), and in other translation directions. Indeed, Paidiom could be easily adapted to other language pairs, since it would only require changing the pretrained statistical model from `es_core_news_sm` (Spanish) to another model available in Spacy, depending on the desired source language and then design the detection and conversion patterns adapted to idioms in the new translation direction.

In addition, the present study could also constitute the basis for further research to assess the escalation and integration of this model into other language-dependent text-preprocessing systems for the automatic translemmatisation of MWEs, with the purpose of enhancing MWE-aware NMT systems, so that NMT can eventually offer a more suitable quality for all stakeholders (users, translators, researchers, developers, etc.) in terms of idiom translation.

Acknowledgements

This research was carried out within the framework of several research projects (ref. PID2020-112818GB-I00, PDC2021-121220-I00, ProyExcel_00540 y TED2021-129789B-I00) at Universidad de Málaga (Spain) and at Research Institute of Multilingual Language Technologies (IUITLM). It was also funded by a post-doc grant entitled *Ayuda para la recualificación del Sistema Universitario Español 2021-2023 (Modalidad «Margarita Salas»)* at Universidad de Málaga and Université catholique de Louvain (Belgium).

References

- Bargmann, Sascha, and Manfred Sailer. 2018. The syntactic flexibility of semantically non-decomposable idioms. In Manfred Sailer and Stella Markantonatou (eds). *Multiword expressions: Insights from a multi-lingual perspective*. Language Science Press, pages 1–29
- Baziotis, Christos, Prashant Mathur, and Eva Hasler. 2023. Automatic Evaluation and Analysis of Idioms in Neural Machine Translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 3682–3700.
- Belinkov, Yonatan, and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ArXiv*. <https://arxiv.org/abs/1711.02173>
- Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *ArXiv*. <https://arxiv.org/abs/1608.04631>
- Chatzikoumi, Eirini. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2): 137-161. doi:10.1017/S1351324919000469
- Colson, Jean-Pierre. 2019. Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic—and a Tentative Corpus-driven Solution. In Gloria Corpas Pastor, Ruslan Mitkov, Maria Kuilovskaya, and María Araceli Losey León

- (eds). *Proceedings of the Third International Conference EUROPHRAS 2019*. Tradulex, pages 145–156.
- Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43(4): 1–92.
- Corpas Pastor, Gloria. 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In Inés Olza and Elvira Manero (eds). *Fraseopragmática*. Berlin: Frank & Timme, pages 335-373.
- Corpas Pastor, Gloria. 2021. Technology Solutions for Interpreters: The VIP System. *Hermēneus. Revista de Traducción e Interpretación*, 23: 91-123.
- DFDEA – Seco, Manuel, Olimpia Andrés, and Gabino Ramos. 2017. *Diccionario fraseológico documentado del español actual, locuciones y modismos españoles (2ª edición)*. Madrid: Aguilar
- DLE – Real Academia Española (n. d.). *Diccionario de la Lengua Española*. <http://www.rae.es>
- Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1): 61–103.
- Hidalgo-Tertero, Carlos Manuel. 2020. Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. In Pedro Mogorrón Huerta (eds). *Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting. MonTI Special Issue 6*: 154-177. <https://doi.org/10.6035/MonTI.2020.ne6.5>
- Hidalgo-Tertero, Carlos Manuel. 2021. El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). In Gloria Corpas Pastor, María del Rosario Bautista Zambrana, and Carlos Manuel Hidalgo-Tertero (eds). *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus*. Granada: Comares, pages 253-270.
- Hidalgo-Tertero, Carlos Manuel. 2024/forthcoming. ¿DeepL, Google Translate o VIP? Qué sistema ofrece un mejor rendimiento en la traducción de locuciones continuas y discontinuas. In Gloria Corpas Pastor, and Francisco Javier Veredas Navarro (eds.). *Tecnologías lingüísticas multilingües: desarrollos actuales y transición digital*. Granada: Comares.
- Hidalgo-Tertero, Carlos Manuel, and Gloria Corpas Pastor. 2020. Bridging the ‘gApp’: improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology*, 11: 61-80. <https://doi.org/10.1515/phras-2020-0005>
- Hidalgo-Tertero, Carlos Manuel, and Gloria Corpas Pastor. 2021. La variación fraseológica: análisis del rendimiento de los corpus monolingües como recursos de traducción. *Études romanes de Brno*, 42(1): 359-379. <https://doi.org/10.5817/ERB2021-1-19>
- Hidalgo-Tertero, Carlos Manuel, and Gloria Corpas Pastor. 2024a/forthcoming. Qué se traerá gApp entre manos... O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). In Míriam Seghiri, and Míriam Pérez Carrasco (eds). *Nuevas tendencias en traducción e interpretación especializadas*. Frankfurt am Main: Peter Lang.

- Hidalgo-Ternero, Carlos Manuel, and Gloria Corpas Pastor. 2024b/forthcoming. ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems. In Johanna Monti, Gloria Corpas Pastor, Ruslan Mitkov, and Carlos Manuel Hidalgo-Ternero (eds). *Recent Advances in MWU in Machine Translation and Translation technology*. Amsterdam & Philadelphia: John Benjamins Publishing Company.
- Hidalgo-Ternero, Carlos Manuel, and Xiaoqing Zhou-Lian. 2022. Reassessing gApp: does MWE discontinuity always pose a challenge to Neural Machine Translation? In Gloria Corpas Pastor, and Ruslan Mitkov (eds). *Computational and Corpus-Based Phraseology*. Cham: Springer, pages 116–132.
- Honnibal, Matthew, and Inés Montani. 2017. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
- Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *Arxiv*. <https://arxiv.org/pdf/1610.01108.pdf>
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2004. *The Sketch Engine*. <https://www.sketchengine.eu>
- Lohar, Pintu, Maja Popovic, Haithem Afli, and Andy Way. 2019. A systematic comparison between SMT and NMT on translating user-generated content. In *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*.
- Martínez Alonso, Héctor, and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural [S.l.]*, 57: 91-98. ISSN 1989-7553.
- Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2017. Learning multilingual named entity recognition from Wikipedia. *figshare*. Dataset. <https://doi.org/10.6084/m9.figshare.5462500.v1>
- Parra Escartín, Carla, Almudena Nevado Llopis, and Eoghan Sánchez Martínez. 2018. Spanish multiword expressions: Looking for a taxonomy. In Manfred Sailer and Stella Markantonatou (eds). *Multiword expressions: Insights from a multi-lingual perspective*. Language Science Press, pages 271–323.
- Ramisch, Carlos, and Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslan Mitkov (ed.). *Oxford Handbook of Computational Linguistics* (2nd ed). Oxford: Oxford University Press.
- Ramisch, Carlos. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Cham: Springer. ISBN 978-3-319-09206-5.
- Rohanian, Omid, Shiva Taslimipour, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1*, pages 2692-2698.
- Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh (ed.). *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science*. Verlin, Heidelberg & Dordrecht: Springer, pages 1–15.

- Santoyo, Julio César, and Rosa Rabadán. 1991. Basic Spanish Terminology for Translation Studies: a Proposal. *Meta*, 36(1): 318-322.
- Wang, Xing, Zhaopeng Tu, Deyi Xiong, and Min Zhang. 2017. Translating phrases in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 1421–1431. DOI: 10.18653/v1/D17-1149
- Zaninello, Andrea, and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3816–3825

Correcting biased translations with the Fairslator API

Michal Měchura

Fairslator

michmech@lexiconista.com

Abstract

This paper introduces the Fairslator API, a software solution for gender rewriting and form-of-address rewriting of translations. Starting with a review of bias (including but not limited to gender-bias) in machine translation and with a brief introduction to the concept of rewriting as a method for solving the problem, the paper then demonstrates how the Fairslator API can be used to rewrite biased translations into alternative genders and forms of address, and surveys the ways in which this technology can be integrated into the translation workflow, for example as a step in machine translation post-editing.

1 Introduction: bias and rewriting

Gender bias is a well-known complication in machine translation (Savoldi *et al.* 2021). A machine-produced translation, however fluent or grammatically correct, may be biased if the machine has made an unjustified assumption about someone’s gender, for example if it has translated a gender-neutral source word into a stereotypically gendered target word (‘pilot’ → ‘male pilot’). This may be contrary to the intended meaning and may require manual post-editing.

A similar problem affects forms of address such as second-person pronouns, especially when translating from English (‘you’) into languages with richer pronoun systems (German ‘du/Sie/ihr’, French ‘tu/vous’). Again, manual post-editing is required to correct these errors.

Recently, semi-automatic rewriting has emerged (Alhafni *et al.* 2022, Moryossef *et al.* 2019, Měchura 2022a) as an alternative to fully manual post-editing. A rewriter is a software tool which takes the output of machine translation as its input and, along with additional instructions about the genders and other properties of people mentioned in the text, rewrites the translation accordingly. Rewriting requires a human to decide, for example, that ‘pilot’ should be interpreted as female or that ‘you’ should be interpreted as plural. Once the human has made these decisions (and communicated them to the tool through menus etc.) the rest of the rewriting process is automatic.

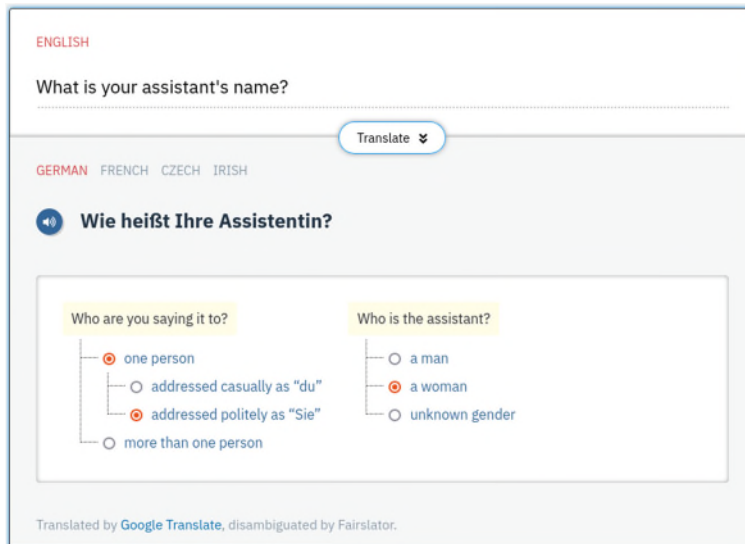


Figure 1. Fairslator

One implementation of the rewriting idea is Fairslator¹ (Měchura 2022b), a tool which rewrites machine translation output in four language pairs (from English into German, French, Czech and Irish). Fairslator scans the pair of texts (the source text plus its translation) for any bias-causing ambiguities in gender and forms of address, asks the user for a disambiguation, and rewrites the translation accordingly. Figure 1 shows Fairslator’s web interface which has been available since 2022.

2 Rewriting biased translations

As of 2023 the functionality of Fairslator is also available through a REST API² which takes input and produces output in a machine-readable, JSON-encoded form. We will now demonstrate how the API works through several examples.

2.1 Example 1: gender rewriting (first person)

An example of input which a software client can send to the Fairslator API is shown in (1). This asks the API to take the French translation of the English source text (which the client has obtained from a third-party machine translation provider) and to rewrite it so that the first person (the person saying the sentence, the ‘I’ and ‘me’ of the sentence) become female: this is set using the `firstPerson` field where the possible values are `m` for male and `f` for female. The output is shown in (2) where the translation of ‘happy’, which is required in French to agree with the subject in gender, has been changed from male *hereux* to female *hereuse*.

¹ <https://www.fairslator.com>

² <https://rapidapi.com/lexiconista/api/fairslator>

(1) First-person gender rewriting: input

```
{
  "sourceText": "I am happy to be here.",
  "sourceLang": "en",
  "text": "Je suis heureux d'être ici.",
  "lang": "fr",
  "firstPerson": "f"
}
```

(2) First-person gender rewriting: output

```
{
  "success": true,
  "originalText": "Je suis heureux d'être ici.",
  "firstPerson": "f",
  "rewrittenText": "Je suis heureuse d'être ici."
}
```

2.2 Example 2: gender rewriting (third person)

A more complicated example is shown in (3). Here, a client application is asking the Fairslator API to take the German translation of the English text and to rewrite it in such a way that the translation of 'nurse' is changed from female to male and the translation of 'patient' from male to female. These instructions are specified in the `thirdPersons` field, the possible values are again `m` and `f`. The output is shown in (4). As you can see, the gender of the nouns have been changed accordingly as have the definite articles that precede them.

(3) Third-person gender rewriting: input

```
{
  "sourceText": "The nurse saved the patient's life",
  "sourceLang": "en",
  "text": "Die Krankenschwester rettete dem Patienten das Leben.",
  "lang": "de",
  "thirdPersons": {
    "nurse": "m",
    "patient": "f"
  }
}
```

(4) Third-person gender rewriting: output

```
{
  "success": true,
  "originalText": "Die Krankenschwester rettete dem Patienten das Leben.",
  "thirdPersons": {
    "nurse": "m",
    "patient": "f"
  },
  "rewrittenText": "Der Krankenpfleger rettete der Patientin das Leben."
}
```

2.3 Example 3: rewriting into gender-neutral forms

In addition to the values m and f, a third value b is possible which causes Fairslator to produce gender-neutral output using various patterns of intra-word punctuation that have been becoming common in many European languages recently. An example is shown in (5) where a client is asking the Fairslator API to rewrite the German translation such that the translation of ‘students’ is changed from the default male to gender-neutral. The output is shown in (6). In German, Fairslator uses the colon (:) to produce gender-neutral rewrites. In French, the middle dot (·) is used: an example is shown in (7).

(5) Gender-neutral rewriting: input

```
{
  "sourceText": "All students must register.",
  "sourceLang": "en",
  "text": "Alle Studenten müssen sich anmelden.",
  "lang": "de",
  "thirdPersons": {
    "students": "b"
  }
}
```

(6) Gender-neutral rewriting: output (German)

```
{
  "success": true,
  "originalText": "Alle Studenten müssen sich anmelden.",
  "thirdPersons": {
    "students": "b"
  },
  "rewrittenText": "Alle Student:innen müssen sich anmelden."
}
```

(7) Gender-neutral rewriting: output (French)

```
{
  "success": true,
  "originalText": "Tous les étudiants doivent s'inscrire.",
  "thirdPersons": {
    "students": "b"
  },
  "rewrittenText": "Tous les étudiant·es doivent s'inscrire."
}
```

2.4 Example 4: form-of-address rewriting

Turning from gender to forms of address, example (8) shows how a client can ask the Fairslator API to take a German translation – which addresses the reader using the informal singular version of ‘you’, *du* – and rewrite it into the formal version of ‘you’, *Sie*. The output is shown in (9).

(8) Form-of-address rewriting: input

```
{
  "sourceText": "Have you remembered it?",
  "sourceLang": "en",
  "text": "Hast du es dir gemerkt?",
  "lang": "de",
  "secondPerson": "v"
}
```

(9) Form-of-address rewriting: output (German)

```
{
  "success": true,
  "originalText": "Hast du es dir gemerkt?",
  "secondPerson": "v",
  "rewrittenText": "Haben Sie es sich gemerkt?"
}
```

The possible values for the `secondPerson` parameter are any combination of s (singular) or f (plural) or nothing, followed by t (informal) or v (formal) or nothing. In German, the values typically used are:

st singular informal, *du*

pt plural informal, *ihr*

v formal, *Sie*

In French, the values typically used are:

st singular informal, *tu*

v singular formal or plural, *vous*

2.5 Example 5: rewriting gender and form-of-address together

Finally, example (10) shows how a client might ask the Fairslator API to rewrite both the gender and the form of address of the same person at the same time. The French translation in the input uses the singular informal form of address (st) and addresses the reader as male. The client is asking to rewrite this into female and formal: notice how any form-of-address code (st, v etc.) can be combined with any gender code (m, f) to give e.g. stm or vf. The output is in (11).

(10) Form-of-address together with gender: input

```
{
  "sourceText": "Are you a good student?",
  "sourceLang": "en",
  "text": "Es tu un bon étudiant?",
  "lang": "fr",
  "secondPerson": "vf"
}
```

(11) Form-of-address together with gender: output

```
{
  "success": true,
  "originalText": "Es tu un bon étudiant?",
  "secondPerson": "vf",
  "rewrittenText": "Êtes vous une bonne étudiante?"
}
```

3 Analyzing biased translations

In addition to the rewriting capabilities shown in the five examples above, the Fairslator API offers an endpoint for analyzing the genders and forms of address of any persons mentioned in the translation, without changing them. Client applications can use this to detect and reveal any biases present in a text. Example (12) shows how a client might ask the API to analyze a German translation and (13) shows the output.

(12) Analyzing a translation: input

```
{
  "sourceText": "Your students like you.",
  "sourceLang": "en",
  "text": "Deine Studenten mögen dich.",
  "lang": "de"
}
```

(13) Analyzing a translation: output

```
{
  "success": true,
  "sourceText": "Your students like you.",
  "text": "Deine Studenten mögen dich.",
  "secondPerson": {
    "registerFreedom": "tv",
    "register": "t",
    "numberFreedom": "sp",
    "number": "s"
  },
  "thirdPersons": [{
    "keyword": "students",
    "genderFreedom": "mfb",
    "gender": "m",
    "number": "p",
  }]
}
```

The field `secondPerson` in the output says that the second person is present in the text: the text addresses the reader.

The line `"registerFreedom": "tv"` says that, in the English original, the form of address is ambiguous in terms of register: it affords the freedom to be interpreted as either `t` (informal) or `v` (formal). The line `"register": "t"` says that out of these two options the German translation uses the `t` option: it addresses the reader informally.

The line "numberFreedom": "sp" says that, in the English original, the form of address is ambiguous in terms of number: it affords the freedom to be interpreted as either s (singular) or p (plural). The line "number": "s" says that out of these two options the German translation uses the s option: it addresses the reader as one person.

The field thirdPersons in the output provides similar facts about any third persons mentioned in the text. In this example, only one third person is present, represented by the English word 'students'.

The line "genderFreedom": "mfb" says that the English word, as used in this sentence, is ambiguous in terms of gender: it affords the freedom to be interpreted as either m (male) or f (female) or b (gender-neutral). The line "gender": "m" says that out of these three options the German translation uses the m option: it refers to the students as males.

The line "number": "p" say that the German translation refers to the students in the plural. There is no numberFreedom field, which means that the word 'students' is not ambiguous in terms of number: it does not afford the freedom to be interpreted in any other way than plural.

The analysis output as exemplified here can be used by an application to “know” about any bias-causing ambiguities in the (English) source text and about any bias in the (German or French) translation.

4 Coverage and accuracy

The Fairslator API currently supports two language pairs: it can analyze and rewrite translations **from English into German** and **from English into French**. Additional language pairs are in development.

In terms of accuracy, no rigorous reproducible tests have been done on Fairslator yet (including the API). Informally, the software has been observed to perform with nearly 90% accuracy on texts of a conversational nature (dialogs, chat transcripts, film subtitles) and with an accuracy of approximately 75% on information-dense factual texts such as Wikipedia articles.

5 Potential uses

One possible use for the Fairslator API is in end-user-facing machine translation applications such as Fairlator's own website which uses the API in two ways. First, Fairslator analyzes each translation and, if the analysis reveals that the translation is biased because the source text contains bias-causing ambiguities, Fairslator gives the user clickable options for resolving the ambiguities manually: it asks whether references to each person should be translated as male or female, whether the reader should be addressed formally or informally, and so on. Then, depending on what the user has selected, Fairslator uses the API again to rewrite the translation accordingly.

It is possible to imagine other uses, such as in machine translation post-editing where editors could make changes to gender and forms of address with one single click instead of typing every change manually, thus automating away some of the more tedious aspects of post-editing. The API could be equally useful in unsupervised scenarios when a tool takes the output of machine translation and uses the API to make sure that, for example, all references to third

persons in the plural are translated as gender-neutral, or that the form of address is informal throughout – all this *before* the translation is made available to a post-editor. And finally, there are potential uses beyond the classical desk-bound translation scenario, such as when translating real-time dialog between two people (adapting genders and forms of address appropriately as people speak) or when customizing user interfaces for specific users.

One way or another, translation rewriting has the potential to become a useful component in the mix of technologies employed during the translation process. This technology is now ready for real-world use.

References

- Alhafni, Bashar, Ossama Obeid, and Nizar Habash. 2022. “The User-Aware Arabic Gender Rewriter.” <https://doi.org/10.48550/ARXIV.2210.07538>.
- Měchura, Michal. 2022a. “We Need to Talk about Bias in Machine Translation: The Fairslator Whitepaper.” <https://www.fairslator.com/fairslator-whitepaper.pdf>.
- Měchura, Michal. 2022b. “Introducing Fairslator: A Machine Translation Bias Removal Tool.” In *Translating and the Computer 44 Proceedings*, 90–95. Luxembourg: Editions Tradulex. <http://www.tradulex.com/varia/TC44-luxembourg2022.pdf>.
- Moryossef, Amit, Roe Aharoni, and Yoav Goldberg. 2019. “Filling Gender & Number Gaps in Neural Machine Translation with Black-Box Context Injection.” In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 49–54. Florence, Italy: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-3807>.
- Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. “Gender Bias in Machine Translation.” *Transactions of the Association for Computational Linguistics* 9 (August): 845–74. https://doi.org/10.1162/tacl_a_00401.

Human & Machine Translation Quality: Comparing & Contrasting Concepts

Bettina Hiebl & Dagmar Gromann

University of Vienna

{bettina.hiebl, dagmar.gromann}@univie.ac.at

Abstract

Quality assurance is a central component of both human and machine translation with different points of view from the perspectives of Translation Studies (TS) and the field of Machine Translation (MT). Whereas TS focuses on the purpose, on pragmatic aspects of translation as well as on comprehensibility, Translation Quality Assessment (TQA) in the field of MT includes TQA frameworks for assessment by humans, consisting of manual error classification, and by automated metrics. In an attempt to bridge the gap between these two fields, this paper focuses on comparing and contrasting central concepts of assessing translation quality in both fields, TS and MT, as well as providing an overview and description of overlapping quality concepts of the two fields, based on an extensive systematic literature review on translation quality (assessment). The detailed descriptions and comparisons of the perspectives from both fields will provide a valuable point of reference for potential intersections of quality concepts in TS and MT.

1 Introduction

Defining and assessing the quality of a translation is a matter of debate in TS, given the multitude of proposed perspectives and approaches on the topic (Koby & Lacruz, 2017). An early idea to merely measure quality by degree of equivalence was soon surpassed. For instance, a highly influential approach, the Skopos theory (Reiss & Vermeer, 1984), focused on preserving the purpose of the source text in the translation. House (2015) even deemed determining the purpose challenging and proposed dividing a text into register and genre with further subdivisions. Others focused on the effects of a translation on the recipients, e.g. Göpferich (2008) introduced comprehensibility dimensions.

While in MT both human and machine translation quality assessment are considered, for decades a strong focus on operationalised definitions and approaches, i.e. automatically measuring the translation quality during training MT models, has prevailed. For instance, the well-known and to this date still applied BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee & Lavie, 2005) represent examples of such automated metrics. To address the shortcomings of these automated approaches, human assessment approaches have been proposed, e.g. Popović (2020).

Translation quality concepts shared by the fields of TS and MT include accuracy and fluency (Castilho et al., 2018). Aside from these shared notions, concepts to define, assess, and measure translation quality differ considerably between TS and MT. A previous survey to tackle this issue concentrated on the expertise and affiliation of publication authors from TS, MT or both (Hiebl & Gromann, 2023). In contrast, this survey seeks to focus on central concepts of translation quality from the perspective of TS, MT, their comparison, and their differences.

2 Method

The methodological basis for this systematic literature review builds on the guidelines by Kitchenham (2004) and the Systematic Reviews and Meta-Analyses (PRISMA) method (Page et al., 2021). The method is presented in line with the three main stages of the PRISMA method: identification, screening and inclusion. Before performing the literature review, a detailed review protocol was drawn up, including, among other things, the main research question or how quality assessment approaches and methods from the two fields can be combined to identify a mutually beneficial, joint quality assessment framework, the search keywords, search platforms as well as inclusion criteria.

2.1 Identification

An initial list of domain-specific keyword combinations was tested on domain-specific search platforms. In TS, targeted journals included Target and Translating and Interpreting Studies; in MT, the journals Machine Translation and TACL as well as the ACL proceedings were used. The final set of 12 keywords and keyword combinations identified was: “human translation” / “machine translation” AND “quality assessment” / “quality estimation” / “quality”; “translation quality”; “translation quality” AND “accuracy” / “assessment” / “comprehensibility” / “estimation” / “fluency”. The search was performed on three major scholarly platforms, i.e., Google Scholar, Web of Science, and Scopus in a search period from 2012 to 2022. This period was selected to include recent literature, but also the move from Statistical MT to Neural MT; important work outside this period was included via snowballing. The resulting publications were ranked based on a keyword score. To this end, two domain experts rated the keyword combinations on a scale from 1 (least important) to 10 (most important), where the final keyword score represented the average of the two scores. This keyword score was multiplied by the times a publication was found, based on the same keyword combination, on different platforms, adding occurrences across keywords and platforms.

2.2 Screening

Duplicates in the final result set were removed and the remaining set was ranked as described in Section 2.1. Starting from the top-ranked ones, publications were screened for their relevance to translation quality and categorised into human translation, machine translation or both.

2.3 Inclusion

The most important criteria for inclusion in the final result set was the topic of translation quality, quality control by means of peer reviewing, and English as publication language.

Including publications in other languages in this literature review would have exceeded both the scope of the study and the resources available.

3 Results

The number of records returned from searching with the 12 keyword combinations was 13,762. After removal of duplicates, the keyword-ranking procedure produced results with a maximum score of 167 for the highest-ranked paper. A previously published review of preliminary results for the same result set had a cutoff score of 77, which resulted in a set of 41 publications, and a focus on grouping publications by affiliation and expertise of the authors (Hiebl & Gromann, 2023). This previous analysis showed that the main ideas of authors from the fields of TS and MT still differ slightly, but that quality concepts are converging and cross-field collaboration increasing. In contrast, the cutoff score for this article was determined at 64.5 to include a higher number of relevant papers. While it is possible that this cutoff score leads to the exclusion of specific quality concepts, in general a high degree of repetition in quality concepts could be observed beyond this score. In the screening process, 7 records were excluded because they were not peer-reviewed, 6 because they were superseded or results were presented elsewhere, and 2 were excluded for being book reviews.

The final result set of 85 publications was extended by performing snowballing, i.e., following up on important references outside the result set, whereby an additional 17 publications were added. These important references were either cited in more than one of the publications included in the result set or were a fundamental basis for these publications. A search for publications published after the original search period did not lead to the inclusion of additional publications. The final result set of 102 publications was subsequently divided into different thematic groups based on the fields addressed: human translation (15 publications), machine translation (63 publications) or both (24 publications). The detailed distribution across thematic subgroups is shown in Fig. 1.

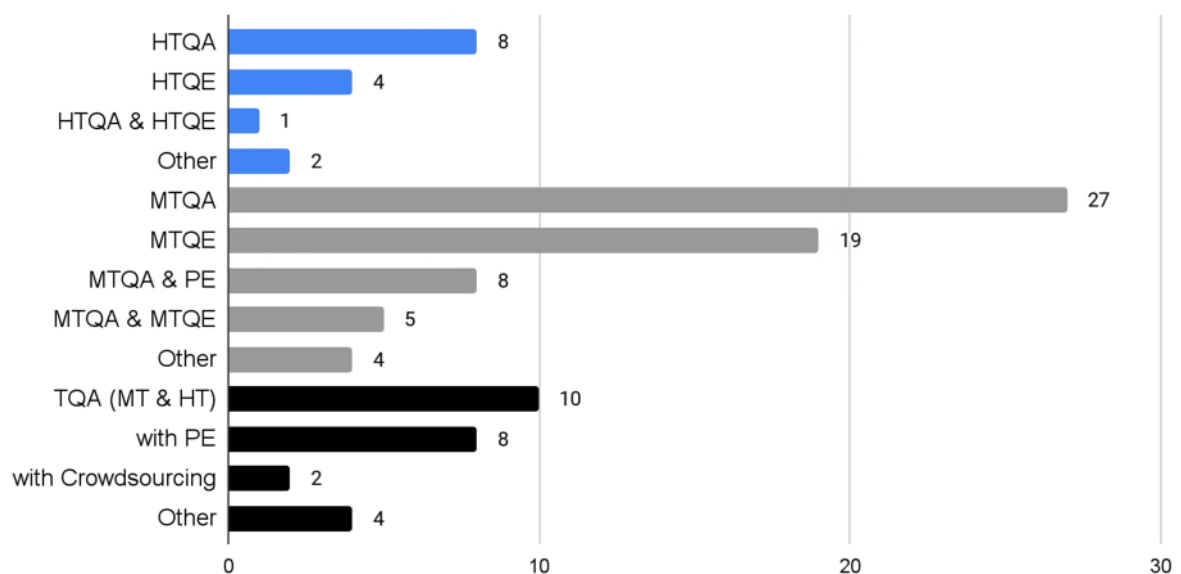


Figure 1: Distribution of result set on specific topics

3.1 Human Translation

The 15 publications in the field of human translation were divided into (i) human translation quality (assessment) HTQA, and (ii) human translation quality estimation HTQE.

HT - Human Translation Quality (Assessment)

There are eight publications on this topic, three of them theoretical and added via snowballing. Reiss and Vermeer (1984) propose a focus on preserving the purpose of the source text in a translation (Skopos theory), while House (2015) focuses more on the register and genre of a text. Göpferich (2008) introduces the so-called dimensions of comprehensibility (concision, correctness, motivation, structure, simplicity, and perceptibility). Leiva Rojo (2018) assesses phraseological quality assuming it correlates with overall quality, however, in most cases the former turned out to be better than the latter.

The remaining four publications use automated metrics or specific tools. Yang et al. (2017) assess students' translations using the quality assessment functionality of a CAT tool as a structured and consistent assessment method, whereas Qassem (2020) studies the student's ability to meaningfully segment cultural references to translation units, recommending training with regards to cultural information as well as the use of technological tools, e.g. eye tracking, Translog, etc. Betanzos et al. (2017) created a corpus of HTs, evaluations, and revisions from a collaborative setting, which they use to train a model that automatically provides feedback on translations, providing, however, very low results and no rigorous model evaluation. Karami et al. (2020) conducted an empirical study on the use of automated metrics to evaluate human translation to test whether a higher number of translations increases the scores, which was only partially confirmed.

HT - Human Translation Quality Estimation

The four publications on this topic show a clear influence from the area of MT and computational linguistics. They propose an evaluation framework based on feature sets extracted from, and used to, assess human translations, focusing on predicting adequacy and fluency (Yuan et al. 2016; 2017). Yuan and Sharoff (2018) assess the influence of Bilingual Multi-Word Units (BMWUs) on trainee translation quality and show that normalised BMWU ratios can be useful for HTQE. The same authors (Yuan & Sharoff, 2020) compare sentence-level HTQE in neural and statistical machine learning approaches, where the former outperform the latter.

Yuan et al. (2018) bring HTQA and HTQE together by analysing the use of cross-lingual terminology extraction to perform terminology-based translation evaluation, where low-frequency terms and term variations remain a challenge. Secară (2005) provides a broad

overview of error-based translation evaluation approaches, from norms, such as SAE J2450, to schemes, such as by the American Translators Association (ATA), to tools for automated error annotation. Nishio and Sutcliffe (2016) suggest that personality traits contribute to making a person a good translator, finding that “an interest in going to the opera, playing scrabble or contract bridge, or enjoyment of cryptic crossword puzzles” (Nishio and Sutcliffe, 2016, p. 63f) helps.

3.2 Machine Translation

The 63 publications in the field of MT can mainly be assigned to (i) machine translation quality (assessment), (ii) machine translation quality estimation, (iii) machine translation quality assessment and post-editing, and (iv) machine translation quality assessment and estimation.

MT - Machine Translation Quality (Assessment)

There are 27 publications on Machine Translation Quality Assessment (MTQA). By snowballing, the automatic metrics BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee & Lavie, 2005), TER and HTER (Snover et al., 2006) were added. Koehn (2009) provides an overview of different methods of MTQA and Popović (2020) proposes a method for manual assessment of MT output without any score, by marking parts of text.

A third of the publications in this category focus on manual assessment of MT. The Multidimensional Quality Metrics (MQM) (Lommel et al., 2013) and TAUS Dynamic Quality Framework (DQF) (Görög, 2014) models, as well as the joint MQM-DQF model (Lommel, 2018), feature prominently in three publications. Burchardt et al. (2016) focus on MT quality in the context of Audio-Visual Translation (AVT), proposing to extend the MQM for AVT-specific categories, such as mistranslations in situative contexts and timing for translations presented out of synch. Candel-Mora (2022) relies on the DQF and proposes to introduce different quality rating scales for different types of texts. Foradi et al. (2022) assess the performance quality of Google Translate from English to Persian and Persian to English using the MQM-DQF model. The MQM keeps on developing by means of a corresponding W3C community group¹.

Also focusing on manual MTQA, Navrátil et al. (2012) compare different syntactic reordering methods for English-German SMT. Graham et al. (2017) assess a methodology for crowdsourcing human MTQA, concluding that evaluation of MT systems by the crowd alone is possible, whereas Fomicheva and Specia (2016) assume that performing MTQA with reference translations may negatively bias human annotators. They show that monolingual evaluation is influenced by the reference provided.

Moorkens (2018) compared an SMT and NMT approach with two cohorts of students using the categories adequacy, post-editing productivity, and error taxonomy, revealing a high preference for NMT. Fonteyne et al. (2020) analysed NMT outputs based on the error categories mistranslation, coherence, style and register, where in this literary genre

¹ <https://www.w3.org/community/mqmcg/>

mistranslations related to accuracy was the biggest source of errors. To improve evaluations, Licht et al. (2022) presented the Cross-Lingual Semantic Textual Similarity (XSTS) metric, which proposes to evaluate five levels of semantic equivalence, from full to none on sentence-level, with emphasis on adequacy rather than fluency. Munkova and Munk (2016), Benkova et al. (2021) and Wang and Ma (2021) focus on automatic metrics for MTQA. While the first focuses on using automatic metrics such as PER, WER and CDER for translation from a minority language (Slovak) and the second compares the assessment of English–Slovak NMT and SMT using BLEU, the third assesses selected works of Xiaoping Deng using the so-called (proposed) digital humanistic method based on a lexical analysis of MT output.

Five publications focus on HTQA and automatic TQA. Popović (2018) provides a theoretical overview of human assessment and automated metrics; Chatzikoumi (2020) does the same on a smaller scale. Rossi and Wiggins (2013) describe the use of HTQA and automated metrics in patent translation, where automated metrics are usually used only as an internal development tool. Toral and Way (2018) compare SMT and NMT systems on novels, performing assessment with automatic metrics, mostly BLEU, as well as human assessment, finding NMT to perform better. Burchardt et al. (2021) argue that different purposes and user groups require different TQA methods and propose the following use cases: (i) a semi-automated method based on regular expressions, (ii) applying MQM, and (iii) a task-based user evaluation.

Four publications focus on humans and MT. Gaspari et al. (2015) conducted a survey of MT competences with 438 respondents, freelance translators and academics, indicating a general increase in MT in translators' workflows. Way (2018) discusses the quality expectations of MT, the possibility of MT enhancing human translation productivity, while emphasising that humans will remain crucial for translation quality. He (2021) concludes that MT can be beneficial for learners. Krüger (2022) seeks to provide input from TS to MT by suggesting that reference translations should be human-approved, contextual factors need to be considered, errors need to be weighted by severity and an evaluation of the added value of MT in professional settings is required.

MT - Machine Translation Quality Estimation

19 publications were attributed to this category. A general overview is provided by Specia and Shah (2018). González-Rubio et al. (2013) present initial methods, and Graham (2015) focuses on a Pearson correlation of gold and prediction distributions.

Most of the publications test and introduce systems with novel features, eight focusing on QE at a word- or sentence-level. Biçici et al. (2013) propose a sentence-level, language-independent method for SMT. Tezcan et al. (2015) present a system to experiment with different learning methods using ensemble methods for word-level QE and single-feature, word-level predictions for sentence-level QE. Aharoni et al. (2014) show that automatic sentence level detection of machine translated texts from monolingual corpora is possible and propose a MTQE technique based on that approach. Taking into account the context, Shah et al. (2015) present new features for MTQE, which are learnt with a continuous space language model. Huang et al. (2020a) propose MTQE based on pre-trained neural language models for sentence- and word-level QE, Ren (2022) for recurrent neural network (RNN)-based sentence-

level methods, and Li et al. (2021) a method based on reinforcement learning. Tingting and Mengyu (2020) combine sentence vectors with RNN vectors to correlate human and machine translation evaluation.

Some less recent, yet interesting approaches, include multi-task and document-level evaluations. De Souza et al. (2014) investigate different multitask learning methods to overcome issues of methods adopting training and test data from different domains. In a similar fashion, De Souza et al. (2015) combine two supervised statistical machine learning paradigms to show that using a single QE component for CAT-tools it is possible to cover multiple translation jobs with different domains and users. Scarton and Specia (2014) focus on document-level QE using discursive features and exploiting pseudo-reference translations. Chen et al. (2021) focus on the document-level and take the context into account, presenting a model based on centering theory.

Liu et al. (2017) propose TQE using only bilingual corpora for word-level MTQE. Chen et al. (2017) propose a neural network and cross-entropy features of source sentences and machine translations to improve language-independent MTQE. The focus of Elmakias and Vilenchik (2021) is on “oblivious MTQE”, meaning that the algorithm does not have access to human judgement scores or the test text’s distribution; this is based on a notion of sentence cohesiveness. In a similar direction, Huang et al. (2020b) assess QE in an unsupervised manner in a black-box setting, without relying on human-annotated data or model-related features.

MT - Machine Translation Quality Assessment & Post-Editing

Only two of the eight publications on making MT output more accessible and acceptable provide a theoretical overview. Han et al. (2021) present an overview of human and automated methods of MTQA, suggesting that future TQA models should involve deeper linguistic features. Maučec and Donaj (2019) focus on integrating human and machine translation, which they see as promising. One of their main points is that MT can serve as a tool to increase translation productivity.

The remaining six publications are empirical studies. While Koponen and Salmi (2015) tested how well a machine-translated text alone can convey the meaning to the reader, working with English to Finnish, and found that approximately half of the time participants were successful in deducing the correct meaning, Castilho and O’Brien (2017) conducted a study assessing the acceptability of MT output among end-users, finding that usability and satisfaction of light post-editing were higher than for the MT output. Sanchez-Torron and Koehn (2016) assessed how different MT systems affect the post-editing process and the product of professional English to Spanish translators, finding that the better the MT quality the less time needed for post-editing and that better MT output leads to better post-editing quality. In their study with translation experts from the German department of the European Commission’s Directorate-General for Translation (DGT), Vardaro et al. (2019) assess how error categories in NMT texts and their post-edited versions are identified and corrected, showing that the most common error types are lexical errors. In their assessment of the ILA Speech-To-Speech (S2S) app, Omazić and Lekić (2021) performed MTQA on different levels:

fluency/adequacy metrics, light post-editing and automated MT evaluation with BLEU, confirming that human and automated assessments correlate and that the translations (English-German and English-Croatian) are of relatively high quality. Their fluency/adequacy (human) assessment uses a framework built on the one proposed by Daems et al. (2014). In this publication, added via snowballing, a fine-grained analysis of MT and PE errors is performed and their relationship is assessed.

MT - Machine Translation Quality Assessment and Estimation

Five of the publications are on the topic of MTQA and MTQE. Lo (2019) presents YiSi, “a unified automatic semantic machine translation quality evaluation and estimation metric” (Lo, 2019, p. 507), which is two-fold: YiSi-1 measures the similarity between a machine translation and human references, YiSi-2 is a reference-less version. Fernandes et al. (2022) propose quality-aware decoding for NMT for MTQA and MTQE. Freitag et al. (2022) focus on minimum Bayes risk decoding, which instead of targeting the hypothesis with the highest probability extracts the hypothesis with the highest estimated quality. In combination with Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) (Sellam et al., 2020), an evaluation metric modelling human judgments based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), it shows significant improvement in human assessment, but the translations are less favoured by surface metrics, such as BLEU. Graham et al. (2016) re-evaluate a couple of human-targeted metrics on a larger scale, evaluating MTQE systems using HTER and Direct Assessment (DA) proposing use of DA for quality estimation evaluations. Rei et al. (2020) was added via snowballing; they present Crosslingual Optimized Metric for Evaluation of Translation (COMET), a neural framework for training multilingual MT evaluation models.

The remaining four publications focus on cognitive effort in post-editing (Vieira, 2014), the results of a systematic review of MT technology in health communication settings (Dew et al., 2018), pre-editing for improving MT quality (Ive et al., 2018) and the description of the creation of a large-scale MT dataset with human annotation, automatically recorded productivity features and manual scoring (Specia et al., 2017).

3.3 Human Translation & Machine Translation

The main topics into which the 24 publications in this area can be divided are (i) translation quality (assessment), (ii) human translation quality (assessment), machine translation quality (assessment) & post-editing, as well as (iii) translation quality (assessment) & crowdsourcing.

HT & MT - Translation Quality (Assessment)

Ten publications can be allocated to translation quality (assessment). Koby and Lacruz (2017) provide an overview of translation and interpreting quality, discussing that the accepted translation quality threshold for machine translated texts is far below the threshold for human translated text, as well as the fact that most TQA metrics expect translation output to be flawed.

Doherty et al. (2018) find that there is mostly no education and training in TQA, which has effects on employability and professional practice of students, which should be changed.

Doherty (2017) focuses on TQA issues from the perspectives of TS, MT, and the translation industry, providing a wish list from the perspective of TS, i.e., more rigour and systematic analysis in TS and more objective results in MT. In the industry, TQA is mostly used to assess translators and is more customer-focused, but there is a lack of agreement on definitions and measurement criteria. Similarly, Castilho et al. (2018) reflect on TQA regarding assessment of human as well as machine translation from different perspectives, i.e., TS, MT and the industry. According to them the main issues regarding TQA are lack of standardisation in its usage, inconsistency, the differing relationship between human and automatic measures, the social quality and risk as well as TQA education and training (Hiebl & Gromann, 2023). An industry perspective on TQA (Marheinecke, 2016) finds that well-defined quality metrics help all stakeholders on the translation markets and that error-annotated MT output will help improve MT quality.

Vela-Valido (2021) describes translation-industry approaches focusing on AI-based translation quality management and the steps performed before, during and after production. From industry, the idea of using NMT and other language model-based approaches to improve the workflow and support humans, where humans have the final decision, are discussed. Back in 2014, Vela et al. (2014) stated that the MT community largely ignored TS, arguing empirically how automated metrics fail to reflect true translation quality, target audience usability, etc.

Lommel et al. (2013) presented the MQM, an adaptable TQA framework for the assessment of human and machine translation. It offers a system of core issue types, e.g. terminology, style, locale conventions, to which different subcategories can be added as needed. Görög (2014) presented a similar framework, the TAUS Dynamic Quality Framework (DQF) with the same objective. These two assessment frameworks were later combined into one, which - among others - is described by Lommel (2018).

HT & MT - Human Translation Quality (Assessment), Machine Translation Quality (Assessment) & Post-Editing

Eight publications are on human and machine TQA and post-editing. In a theoretical approach, Mellinger (2018) calls for re-thinking the concept of translation quality in the digital age, focusing on revision and editing. Increased use of technology in the translation workflow, i.e., MT and CAT, changes the workload distribution, why the definition of translation quality should include compliance with client specifications, the purpose and target audience. Martikainen (2017) presents a functional approach to TQA, namely categorising sources of translational distortion in abstracts of systematic medical reviews.

In addition to the result set, via snowballing there is a publication on a two-step TQA approach, focusing on the dichotomy of adequacy and acceptability (Daems et al., 2013). They test their approach by comparing HT with post-editing and positive results on its usability.

Several other publications compare the quality of HT and post-edited texts. Ortiz-Boix and Matamala (2017) compare post-edited MT to HT from parts of wildlife documentaries, using grading, assessment with MQM, and questionnaires, confirming that there is no significant quality difference between the two categories. Jia et al. (2019) compared translation from-scratch with post-editing of NMT, finding that post-editing was significantly faster with less cognitive effort, and that the fluency and accuracy of post-edited texts were equivalent to those of translated texts.

Carl and Toledo Báez (2019) conducted a study with translators annotating Spanish and Simplified Chinese MT output using an MQM-derived error taxonomy. Assessing the effect of MT errors on post-editing effort they found that accuracy errors influence production and reading duration, and that segments with MT accuracy issues in one language combination are likely to be difficult to translate into other languages (Hiebl & Gromann, 2023).

Munkova et al. (2021b) assess the influence of MT quality on post-editing performance, showing that the translator's performance is influenced by MT quality and that post-editing compared to human translation is more effective. Munkova et al. (2021a) analyse the role of automated evaluation techniques in online professional translator training using residuals of metrics of accuracy (BLEU) and error rate such as PER and WER, finding that these can identify errors in post-editing.

HT & MT - Translation Quality (Assessment) & Crowdsourcing

A topic influenced by recent technological developments is translation quality and crowdsourcing. Whereas traditionally, translation had to be of extremely high quality and only provided by professional translators, some dynamic aspects have been added by the idea of fit-for-purpose translations and crowdsourcing (Jiménez-Crespo, 2017). Different grades of quality entered the market depending on the purpose of the final product, meaning not only the translator is responsible for quality but also the client who decides on the assignment of the translation, i.e., crowd, collaborative translation, individual professional translator, etc. (Jiménez-Crespo, 2018).

The remaining four publications on this topic focus on quality estimation of Arabic to English translations (Ali et al., 2020), a deep learning system which on the sentence-level reaches translation quality comparable to human professionals (Popel et al., 2020), identifying the MT error types with the greatest impact on post-editing effort (Daems et al., 2017), and translationese identification (Rubino et al., 2016).

4 Discussion

The publications included in the result set of the literature review (102) are on the topic of MT (63), MT & HT (24), and on HT (15), which clearly marks the importance of MT in research. As is clearly visible from the analysis, quality assessment as well as quality estimation are the main topics in both MT and HT. In HT, the influence of MT practices is easily discernible

through the use of automatic metrics as well as quality estimation. On the other hand, in MT, the influence of the, originally HT-oriented, translation studies appears as context comes to play an ever more important role in MT, for example when the focus in quality estimation gradually switches from word- or sentence-level QE to document level QE. Machine Translation Post-Editing has grown with the increase in machine translation systems and can be found in publications in the field of MT as well as in HT & MT publications. The publications on this topic range from those assessing the effort of post-editing to those comparing the quality of post-edited machine translated texts with the quality of texts translated by humans, and therefore mark another intersection point between the two disciplines. A concept introduced in both fields is the topic of fit-for-purpose in combination with crowdsourcing, i.e., research focusing on the fact that for different use cases translations of different quality are needed.

Thus, from the above it can be seen that both fields increasingly, mutually influence each other, even though the uptake of TS quality concepts in MT is only partial and limited. One desideratum, also proposed by other authors, would be to include more contextual factors, including culture-specific aspects of the source and target culture or target audience, in MT quality concepts. At the same time, TS research still focuses slightly more on theoretical perspectives and the topic of HT. From the results, it can also be seen that few studies actively include the approach to translation quality from the industry perspective and those that do focus mainly on the utility of NMT and TQA for higher productivity and translation speed. Moreover, in this result set few studies directly associate translation quality with the potential cognitive load and limited creativity of post-editing, an interesting perspective to be further investigated in future studies. Whereas the publications in the result set of this survey focus mainly on results from the research community, their contents are also relevant for translators. The fit-for-purpose concept, for example, is not only an intersection point of MT and TS, but also the industry, focusing on end-user expectations and economic components, price, time, etc. As the above-mentioned topics are of importance to TS, MT, and the industry, a joint focus on them in the future would be beneficial for all fields: TS would benefit from translators' experience and metrics developed by MT specialists, MT would benefit from the vast knowledge of translators and languages as well as context-related ideas of TS and translators' experience, while translators and the industry would benefit from research on translators and languages (from TS) and the assessment and estimation metrics devised by MT.

A number of limitations of this survey must be acknowledged. First of all, there is a fairly limited number of publications included in the result set (102), which raises no claim as to completeness. Secondly, only publications in English were included, which may exclude relevant quality concepts published in other languages. Lastly, focusing on the fields of MT and TS, the number of publications with an industry perspective is limited and an increase would be highly welcome. Nevertheless, the survey and its results provide a good overview of translation quality concepts for MT, TS and industry experts and interested parties as well as some ideas for a closer collaboration.

5 Conclusion and Future Research

The analysis of the literature on translation quality in human and machine translation shows that the field of translation studies still focuses slightly more on theoretical concepts and

machine translation more on metrics, but the mutual influences are clearly visible. Over the last decade, there is a clearly noticeable convergence of the topics and concepts from the two fields; the main topics, i.e., automatic metrics, quality estimation, post-editing and fit-for-purpose, are present in all fields. The results clearly show that TS and MT have learned a lot from each other in recent years; TS has benefited from the technical and computational input from MT, whereas MT has made use of traditional concepts of translations studies, such as context or including humans in the process of machine translation. In future research, the result set should be further enlarged, not only by adding more publications via snowballing, but also by adding more languages and publications. In addition, a TQA criteria catalogue, as opposed to counting errors, combining concepts from the fields of translation studies and machine translation should be devised and evaluated, incorporating the translation industry and professional translators.

References

- Aharoni, R., Koppel, M., & Goldberg, Y. (2014). Automatic detection of machine translated text and translation quality estimation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 289–295.
- Ali, M. S., Alatawi, A., Alsaifi, B., & Noorwali, N. (2020). QUES: A Quality Estimation System of Arabic to English Translation. *International Journal of Advanced Computer Science and Applications*, 11(7).
- Ammann, M. (1990). Anmerkungen zu einer Theorie der Übersetzungskritik und ihrer praktischen Anwendung. *TEXTconTEXT*, 5, 209–250.
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.
- Benkova, L., Munkova, D., Benko, L., & Munk, M. (2021). Evaluation of English–Slovak neural and statistical machine translation. *Applied Sciences*, 11(7), 2948.
- Betanzos, M., Costa-jussà, M. R., & Belanche, L. (2017). TradARES: A tool for the automatic evaluation of human translation quality within a MOOC environment. *Applied Artificial Intelligence*, 31(3), 288–297.
- Biçici, E., Groves, D., & van Genabith, J. (2013). Predicting sentence translation quality using extrinsic and language independent features. *Machine Translation*, 27, 171–192.
- Burchardt, A., Lommel, A., Bywood, L., Harris, K., & Popović, M. (2016). Machine translation quality in an audiovisual context. *Target*, 28(2), 206–221.
- Candel-Mora, M. Á. C. (2022). Fine-tuning machine translation quality-rating scales for new digital genres: The case of user-generated content. *ELUA: Estudios de Lingüística. Universidad de Alicante*, 38, 117–136.
- Carl, M., & Toledo Báez, M. C. (2019). Machine translation errors and the translation process: A study across different languages. *Journal of Specialised Translation*, 31, 107–132.
- Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 9–38). Springer.
- Castilho, S., & O'Brien, S. (2017). Acceptability of machine-translated content: A multi-language evaluation by translators and end-users. *Linguistica Antverpiensia, New Series—Themes in Translation Studies*, 16, 120–136.
- Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2), 137–161.
- Chen, Y., Zhong, E., Tong, Y., Qiu, Y., & Shi, X. (2021). A Document-Level Machine Translation Quality Estimation Model Based on Centering Theory. *Machine Translation: 17th China Conference, CCMT 2021, Xining, China, October 8–10, 2021, Revised Selected*

Papers 17, 1–15.

- Chen, Z., Tan, Y., Zhang, C., Xiang, Q., Zhang, L., Li, M., & Wang, M. (2017). Improving machine translation quality estimation with neural network features. *Proceedings of the Second Conference on Machine Translation*, 551–555.
- Daems, J., Macken, L., & Vandepitte, S. (2014). On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 62–66). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/532_Paper.pdf
- Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+ PE. *Proceedings of the 2nd Workshop on Post-Editing Technology and Practice*, 64–71.
- Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, 8, 1282.
- De Souza, J. G., Negri, M., Ricci, E., & Turchi, M. (2015). Online multitask learning for machine translation quality estimation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 219–228.
- De Souza, J. G., Turchi, M., & Negri, M. (2014). Machine translation quality estimation across domains. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 409–420.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Dew, K. N., Turner, A. M., Choi, Y. K., Bosold, A., & Kirchoff, K. (2018). Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*, 85, 56–67.
- Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, 138–145.
- Doherty, S. (2017). Issues in human and automatic translation quality assessment. In *Human issues in translation technology*. Routledge, London (pp. 131–148).
- Doherty, S., Moorkens, J., Gaspari, F., & Castilho, S. (2018). On education and training in translation quality assessment. In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.),

- Translation Quality Assessment: From Principles to Practice* (pp. 95–106). Springer.
- Elmakias, I., & Vilenchik, D. (2021). An oblivious approach to machine translation quality estimation. *Mathematics*, 9(17), 2090.
- Fernandes, P., Farinhas, A., Rei, R., C. de Souza, J. G., Ogayo, P., Neubig, G., & Martins, A. (2022). Quality-Aware Decoding for Neural Machine Translation. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1396–1412). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.100>
- Fomicheva, M., & Specia, L. (2016). Reference Bias in Monolingual Machine Translation Evaluation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 77–82.
- Fonteyne, M., Tezcan, A., & Macken, L. (2020). Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. *12th International Conference on Language Resources and Evaluation (LREC)*, 3783–3791.
- Foradi, Z., Faroughi, J., & Rezaeian Delouei, M. R. (2022). Assessing the Performance Quality of Google Translate in Translating English and Persian Newspaper Texts Based on the MQM-DQF Model. *Journal of Language and Translation*, 12(4), 107–118.
- Freitag, M., Grangier, D., Tan, Q., & Liang, B. (2022). High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. *Transactions of the Association for Computational Linguistics*, 10, 811–825. https://doi.org/10.1162/tacl_a_00491
- Gaspari, F., Almaghout, H., & Doherty, S. (2015). A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives*, 23(3), 333–358.
- González-Rubio, J., Navarro-Cerdán, J. R., & Casacuberta, F. (2013). Dimensionality reduction methods for machine translation quality estimation. *Machine Translation*, 27, 281–301.
- Göpferich, S. (2008). *Textproduktion im Zeitalter der Globalisierung: Entwicklung einer Didaktik des Wissenstransfers* (3. Aufl.). Stauffenburg.
- Görög, A. (2014). Quality evaluation today: The dynamic quality framework. *Proceedings of Translating and the Computer* 36, 155–164.
- Graham, Y. (2015). Improving Evaluation of Machine Translation Quality Estimation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1804–1813.
- Graham, Y., Baldwin, T., Dowling, M., Eskevich, M., Lynn, T., & Tounsi, L. (2016). Is all that Glitters in Machine Translation Quality Estimation really Gold? In Y. Matsumoto & R.

- Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3124–3134). The COLING 2016 Organizing Committee. <https://aclanthology.org/C16-1294>
- Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1), 3–30.
- Han, L., Smeaton, A., & Jones, G. (2021). Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, 15–33.
- He, X. (2021). Evaluation of Machine Translation Quality Based on Neural Network and Its Application on Foreign Language Education. *AIAM2021: 3rd International Conference on Artificial Intelligence and Advanced Manufacture*, 1395–1399.
- Hiebl, B., & Gromann, D. (2023). Quality in Human and Machine Translation: An Interdisciplinary Survey. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 375–384.
- House, J. (2015). *Translation quality assessment: Past and present*. Routledge.
- Huang, H., Di, H., Xu, J., Ouchi, K., & Chen, Y. (2020a). Ensemble Distilling Pretrained Language Models for Machine Translation Quality Estimation. *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, 231–243.
- Huang, H., Di, H., Xu, J., Ouchi, K., & Chen, Y. (2020b). Unsupervised machine translation quality estimation in black-box setting. *Machine Translation: 16th China Conference, CCMT 2020, Hohhot, China, October 10-12, 2020, Revised Selected Papers 16*, 24–36.
- Ive, J., Max, A., & Yvon, F. (2018). Reassessing the proper place of man and machine in translation: A pre-translation scenario. *Machine Translation*, 32(4), 279–308.
- Jia, Y., Carl, M., & Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *The Journal of Specialised Translation*, 31(1), 60–86.
- Jiménez-Crespo, M. A. (2017). How much would you like to pay? Reframing and expanding the notion of translation quality through crowdsourcing and volunteer approaches. *Perspectives*, 25(3), 478–491.
- Jiménez-Crespo, M. A. (2018). Crowdsourcing and translation quality: Novel approaches in the language industry and translation studies. In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 69–93). Springer.
- Karami, S., Nejadansari, D., & Hesabi, A. (2020). Reliability of Human Translations' Scores Using Automated Translation Quality Evaluation Understudy Metrics. *Journal of Foreign Language Research*, 10(3), 618–629.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004), 1–26.

- Koby, G. S., & Lacruz, I. (2017). The thorny problem of translation and interpreting quality. *Linguistica Antverpiensia, New Series—Themes in Translation Studies*, 16, 1–12.
- Koehn, P. (2009). *Statistical Machine Translation* (1st ed.). Cambridge University Press.
- Koponen, M., & Salmi, L. (2015). On the correctness of machine translation: A machine translation post-editing task. *The Journal of Specialised Translation*, 23(23), 118–136.
- Krüger, R. (2022). Some Translation Studies informed suggestions for further balancing methodologies for machine translation quality evaluation. *Translation Spaces*, 11(2), 213–233.
- Leiva Rojo, J. (2018). Phraseology as indicator for translation quality assessment of museum texts: A corpus-based analysis. *Cogent Arts & Humanities*, 5(1), 1442116.
- Li, F., Zhao, Y., Yang, F., & Cui, R. (2021). Incorporating translation quality estimation into chinese-korean neural machine translation. *Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings*, 45–57.
- Licht, D., Gao, C., Lam, J., Guzman, F., Diab, M., & Koehn, P. (2022). Consistent Human Evaluation of Machine Translation across Language Pairs. *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 309–321.
- Liu, L., Fujita, A., Utiyama, M., Finch, A., & Sumita, E. (2017). Translation quality estimation using only bilingual corpora. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9), 1762–1772.
- Lo, C. (2019). YiSi—A Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névél, M. Neves, M. Post, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 507–513). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-5358>
- Lommel, A. (2018). Metrics for translation quality assessment: A case for standardising error typologies. In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 109–127). Springer.
- Lommel, A., Burchardt, A., & Uszkoreit, H. (2013). Multidimensional quality metrics: A flexible system for assessing translation quality. *Proceedings of Translating and the Computer* 35.
- Marheinecke, K. (2016). Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective. *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, 71–75.
- Martikainen, H. (2017). A functional approach to translation quality assessment: Categorizing sources of translational distortion in medical abstracts. *Linguistica Antverpiensia, New Series—Themes in Translation Studies*, 16, 106–121.

- Maučec, M. S., & Donaj, G. (2019). Machine translation and the evaluation of its quality. *Recent Trends in Computational Intelligence*, 143.
- Mellinger, C. D. (2018). Re-thinking translation quality: Revision in the digital age. *Target*, 30(2), 310–331.
- Moorkens, J. (2018). What to expect from Neural Machine Translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4), 375–387.
- Munkova, D., & Munk, M. (2016). Automatic metrics for machine translation evaluation and minority languages. *Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015: MedCT 2015 Volume 2*, 631–636.
- Munkova, D., Munk, M., Benko, L., & Hajek, P. (2021a). The role of automated evaluation techniques in online professional translator training. *PeerJ Computer Science*, 7, e706.
- Munkova, D., Munk, M., Welnitzova, K., & Jakobovicova, J. (2021b). Product and process analysis of machine translation into the inflectional language. *SAGE Open*, 11(4), 21582440211054501.
- Navrátil, J., Visweswariah, K., & Ramanathan, A. (2012). A comparison of syntactic reordering methods for english-german machine translation. *Proceedings of COLING 2012*, 2043–2058.
- Nishio, N., & Sutcliffe, R. F. (2016). Opera goer or Scrabble player: What makes a good translator? *Machine Translation*, 30(1–2), 63–109.
- Omazić, M., & Lekić, M. (2021). Assessing speech-to-speech translation quality: Case study of the ILA S2S app. *Hieronymus: Journal of Translation Studies and Terminology*, 8, 1–26.
- Ortiz-Boix, C., & Matamala, A. (2017). Assessing the quality of post-edited wildlife documentaries. *Perspectives*, 25(4), 571–593.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., & others. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88, 105906.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.
- Popel, M., Tomkova, M., Tomek, J., Kaiser, \Lukasz, Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020). Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1), 4381.
- Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 129–158). Springer.
- Popović, M. (2020). Informative manual evaluation of machine translation output. *Proceedings*

- of the 28th International Conference on Computational Linguistics, 5059–5069.
- Qassem, M. (2020). Translation unit and quality of translation: Cultural and innovation perspectives. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, 33(2), 536–561.
- Qing Wang & Xiao Ma. (2021). Machine Translation Quality Assessment of Selected Works of Xiaoping Deng Supported by Digital Humanistic Method. *International Journal of Applied Linguistics and Translation*, 7(2), 59–68. <https://doi.org/10.11648/j.ijalt.20210702.15>
- Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.213>
- Reiss, K., & Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie* (Vol. 147). Max Niemeyer Verlag.
- Ren, B. (2022). Machine Automatic Translation Quality Evaluation Model Based on Recurrent Neural Network Algorithm. *Cyber Security Intelligence and Analytics: The 4th International Conference on Cyber Security Intelligence and Analytics (CSIA 2022), Volume 1*, 1019–1026.
- Rossi, L., & Wiggins, D. (2013). Applicability and application of machine translation quality metrics in the patent field. *World Patent Information*, 35(2), 115–125.
- Rubino, R., Lapshinova-Koltunski, E., & Van Genabith, J. (2016). Information density and quality estimation features as translationese indicators for human translation classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 960–970.
- Sanchez-Torron, M., & Koehn, P. (2016). Machine translation quality and post-editor productivity. *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, 16–26.
- Scarton, C., & Specia, L. (2014). Document-level translation quality estimation: Exploring discourse and pseudo-references. *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 101–108.
- Secară, A. (2005). Translation evaluation: A state of the art survey. *Proceedings of the eCoLoRe/MeLLANGE Workshop*, 39–44.
- Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7881–7892). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.704>
- Shah, K., Ng, R. W., Bougares, F., & Specia, L. (2015). Investigating continuous space language models for machine translation quality estimation. *Proceedings of the 2015*

- Conference on Empirical Methods in Natural Language Processing*, 1073–1078.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231.
- Specia, L., Harris, K., Blain, F., Burchardt, A., Macketanz, V., Skadin, I., Negri, M., & Turchi, M. (2017). Translation quality and productivity: A study on rich morphology languages. *Proceedings of Machine Translation Summit XVI: Research Track*, 55–71.
- Specia, L., & Shah, K. (2018). Machine translation quality estimation: Applications and future perspectives. In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 201–235). Springer.
- Tezcan, A., Hoste, V., Desmet, B., & Macken, L. (2015). UGENT-LT3 SCATE system for machine translation quality estimation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 353–360.
- Tingting, L., & Mengyu, X. (2020). Analysis and evaluation on the quality of news text machine translation based on neural network. *Multimedia Tools and Applications*, 79, 17015–17026.
- Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text? In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 263–287).
- Vardaro, J., Schaeffer, M., & Hansen-Schirra, S. (2019). Translation quality and error recognition in professional neural machine translation post-editing. *Informatics*, 6(3), 41.
- Vela, M., Schumann, A.-K., & Wurm, A. (2014). Beyond linguistic equivalence. An empirical study of translation evaluation in a translation learner corpus. *Proceedings of the EACL 2014 Workshop on Humans and Computer-Assisted Translation*, 47–56.
- Vela-Valido, J. (2021). Translation quality management in the AI age. New technologies to perform translation quality management operations. *Revista Tradumàtica*, 93–111.
- Vieira, L. N. (2014). Indices of cognitive effort in machine translation post-editing. *Machine Translation*, 28(3–4), 187–216.
- Way, A. (2018). Quality Expectations of Machine translation. In S. Castilho, S. Doherty, F. Gaspari, & J. Moorkens (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 159–178). Springer.
- Yang, J., Ciobanu, Dragos, Reiss, Caroline, & Secară, Alina. (2017). Using computer assisted translation tools' translation quality assessment functionalities to assess students' translations. *The Language Scholar*, 1, 1–17.
- Yuan, Y., Babych, B., & Sharoff, S. (2017). Reference-free system for automated human translation quality estimation. *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–5.
- Yuan, Y., Gao, Y., Zhang, Y., & Sharoff, S. (2018). Cross-lingual Terminology Extraction for

- Translation Quality Estimation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 3774–3780.
- Yuan, Y., & Sharoff, S. (2020). Sentence Level Human Translation Quality Estimation with Attention-based Neural Networks. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1858–1865.
- Yuan, Y., & Sharoff, S. (2018). Investigating the Influence of Bilingual MWU on Trainee Translation Quality. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1981–1988.
- Yuan, Y., Sharoff, S., & Babych, B. (2016). MoBiL: A Hybrid Feature Set for Automatic Human Translation Quality Assessment. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3663–3670.

Subtitling videos within a language service: a hands-on approach

Samuel Urscheler

Vaudoise Insurance

surscheler@vaudoise.ch

Sandra Casas

Vaudoise Insurance

scasas@vaudoise.ch

Abstract

The internal language service of “Vaudoise Insurance”, a Swiss insurance company, translates, copywrites and proofreads over 1,900 mandates a year. Thanks to its visibility strategy, it successfully positioned itself as the sole purveyor of subtitles.

Before drafting processes for video subtitling, we conducted a benchmark survey of current tools and technologies. We wanted a user-friendly, cheap, easy-to-install desktop tool, if possible, with speech recognition. We settled on the free open-source tool Subtitle Edit.

We follow two processes depending on the type of videos:

1. Videos with scripts only require spotting, revision, and export steps.
2. Videos without scripts require additional speech recognition and translation/spotting steps, followed by revision and export.

For interlingual videos, we translate directly while using the spotting from the intralingual video, then perform the revision and export steps. If there is no intralingual version, we compress translating and spotting into one step. Also, no previous transcription is needed in this case; the content is mostly colloquial and easy to translate.

For our intents and purposes, Subtitle Edit does the trick. The process is optimized and easily understood by our translator colleagues, and clients. Better speech recognition for Swiss German would be a plus in the future.

1 Positioning

The internal language service of the Swiss insurance company “Vaudoise Insurance” translates, copywrites and proofreads over 1,900 mandates a year. As a language service in a Swiss insurance company, we have seen the volume of audiovisual content grow significantly over the past couple of years. As a result, subtitling (“ST”) our company’s own videos has become part of our daily business.

It appeared logical to position ourselves as the provider of subtitles, as we possess the necessary technical and soft skills: task management, CAT tools, localization, cultural/regional differences, or adjustments towards target audiences to name a few. Piggybacking on a previous visibility and renaming campaign for our internal NMT tool and our language service (recent “Language competence center”), we successfully positioned ourselves as ST experts.

2 Benchmark

Before drafting processes for video subtitling, we conducted a benchmark survey of current tools and technologies. We wanted a **user-friendly, cheap, easy-to-install desktop tool**, if possible, with speech recognition and a CAT interface (at least at first). The survey revealed almost no price-accessible ST tools with an MT or CAT interface; the few MT-ST solutions on the market are mostly expensive and cloud-based or outsourced all-in-one solutions. After research, **we settled on Subtitle Edit (SE)**. SE is a free, open-source desktop tool with regular updates, speech recognition (two different systems) and all the necessary features. There are still however some disadvantages, such as no CAT interface, no automatic recognition of shot

changes, almost no formatting in .srt format, though extensive formatting is possible in .ass format.

3 Processes

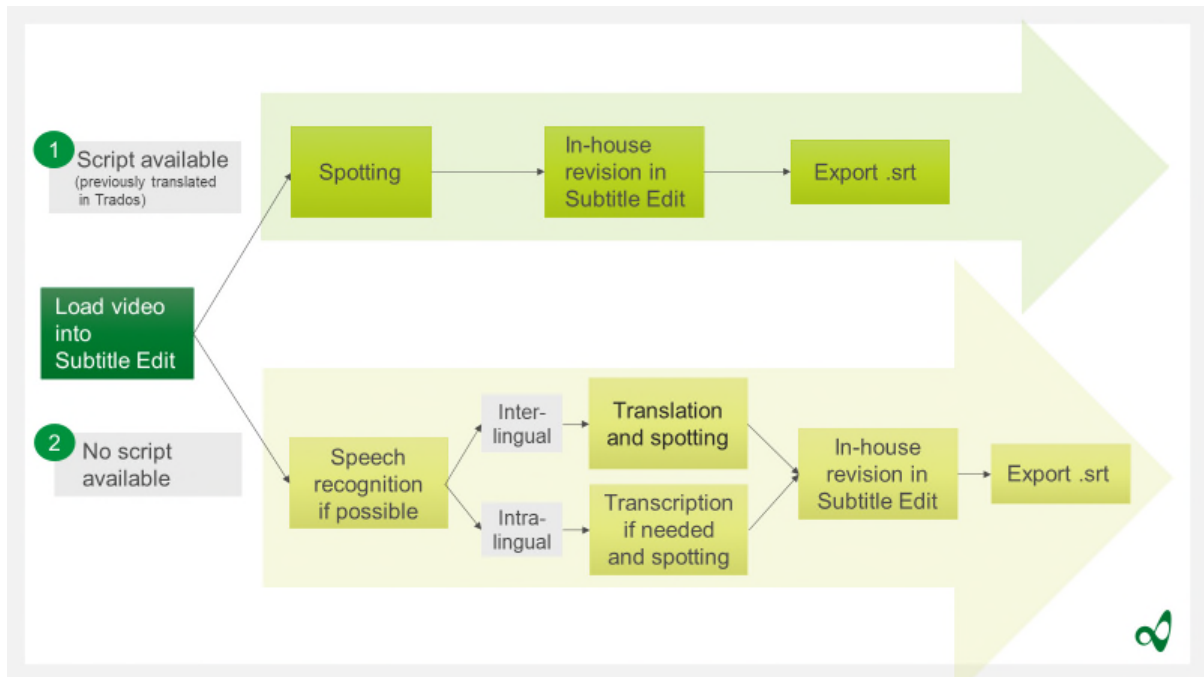


Figure 16. Subtitling processes

3.1 Description

As detailed in the screenshot above, processes for type 1) and type 2) videos only differ in two steps:

- Type 1) videos only need spotting, revision, and export.
- Type 2) videos:
 - intralingual subtitles (i.e., FR audio, FR ST) speech recognition can be applied before transcription and spotting, then revision, and export.
 - interlingual subtitles (e.g., FR audio, DE ST): we copy the intralingual .srt (thus taking over its spotting) and translate it using SE's translation feature).

As seen in the screenshot, for type 2) interlingual videos **we translate directly while using the spotting from the intralingual video**, then do the revision and export steps. If there is no intralingual version, we compress translation and spotting into one step. This illustrates our hands-on approach: internal testing revealed that compressing translation and spotting into one step is faster. Also, no previous transcription of the source language content for subsequent translation is needed in this case; the content is mostly colloquial and easy to translate. After export, our video specialist imports the .srt into the video file and applies corporate design to the ST.

3.2 CAT tool integration

The fact that SE does not have a CAT tool feature did not pose a problem in the end. At Vaudoise, there are mainly two types of videos with requiring ST:

1. info-heavy, speaker(s) reading a script from a teleprompter.
 - a. No CAT interface needed during ST phase, because script is translated in our CAT tool beforehand. We then copy it into SE while doing the spotting.
2. dialogue-heavy, no script, interviews and/or spontaneous speech.
 - a. No CAT interface strictly necessary, given the unique and spoken content of this type of ST. If needed, FR, DE and IT .srt files can still be aligned in our CAT tool afterwards.

Both types can be in DE, FR, or IT, or bilingual/trilingual, and clients mostly request one video version per language, which often leaves us with at least one intralingual version (for example, a video with FR audio and FR ST, a version with DE ST, and another one with IT ST).

4 Successes and room for improvement

We implemented our new ST process a year ago. For our intents and purposes, Subtitle Edit does the trick, we do not need a fancy subtitling tool or an expensive all-in-one solution for now. The process is optimized and easily understood by our translator colleagues, and clients.

Although our current process meets our needs, speech recognition software is still under par, especially for Swiss German which is frequently used in our videos. SE's speech recognition plug-in only features standard German. Some Swiss transcription and subtitling services/projects are attempting to bridge this gap: [töggel](#), [PASSAGE](#) and a research project by [Dr. Vogel and team at ZHAW](#). Some yielded mixed results for our purposes but may be of great use in the future.

From shifting thoughts to unlocking knowledge: The power of terminology in the digital era

Denis Dechandon

Publications Office of the European
Union

denis.dechandon@publications.europa.eu

u

Anikó Gerencsér

Publications Office of the European
Union

aniko.gerencser@publications.europa.eu

Lucy Walhain

Publications Office of the European
Union

lucy.walhain@publications.europa.eu

Carolina Dunaevsky

Court of Justice of the European Union

Carolina.Dunaevsky@curia.europa.eu

Mihai Paunescu

Publications Office of the European Union

mihai.paunescu@ext.publications.europa.eu

u

Abstract

Following the presentation delivered at Translating and the Computer Conference 2023 (TC45) with the same title, this paper explores the transformative impact of semantic technologies on translation and terminology management. It delves into the shift from traditional management of terms to knowledge-driven approaches and how semantic technologies enable improved translation efficiency, accuracy, and interoperability in the age of artificial intelligence (AI).

The paper introduces semantic technologies and their significance in modern terminology practices, highlighting their role in complementing traditional methods through enhanced interoperability and machine readability.

The integral link between semantic technologies and AI is also explored, illustrating how semantic enrichment, knowledge graphs, and ontologies bolster AI-powered tools, elevating their language processing capabilities.

The objective of this paper is to emphasise the power of using Knowledge Organization Systems (KOS) in terminology work and showcase the use of the VocBench collaborative workflow by the European Commission as an example of using semantic technologies in the real world, both to improve terminology processes and to adapt to the AI-driven environment.

1 Introduction

In the dynamic landscape of language and knowledge management, the advent of digitalisation and artificial intelligence (AI) marks a watershed moment, prefiguring an era replete with transformative challenges and opportunities. This discourse delves into the profound impact of semantic technologies on knowledge management and terminology management – a discipline traditionally anchored in the systematic aggregation, structuring and dissemination of terminologies intrinsic to specialised knowledge domains. Amidst the accelerating pace of global communication and the burgeoning need for information exchange, the imperative for precise and efficient terminology management has surged, as an essential linchpin in bridging multifarious languages with the exponentially expanding corpus of human knowledge.

At its essence, terminology management embodies an interdisciplinary crusade, interweaving the strands of linguistics, information science and domain-specific acumen. In our intricately connected, globalised milieu, the import of terminology transcends conventional lexicons or dictionaries. It entails a deep-seated comprehension of concepts and their interrelations, thus enabling lucid and accurate communication across diverse linguistic and cultural landscapes. The metamorphosis of this domain is not merely an academic preoccupation but a pragmatic exigency, pivotal for fostering effective international cooperation, legal precision, scientific breakthroughs and cultural comprehension.

Yet, the swift advancements in AI and the burgeoning of digital platforms present a dual-edged sword for conventional terminology practices. On one flank, AI's ability in language processing, data analytics and pattern recognition augur potent tools for augmenting the breadth and precision of terminology endeavours. Conversely, the fluid nature of language, coupled with the intricacies of human cognition and cultural heterogeneity, poses enduring challenges for AI algorithms. These challenges transcend the technical field, venturing into the conceptual, necessitating a profound grasp of the symbiotic relationship between natural language and knowledge.

The incorporation of semantic technologies into terminology management signifies a pivotal paradigm shift. Semantic technologies, encompassing standards like RDF (Resource Description Framework), OWL (Web Ontology Language) and SKOS (Simple Knowledge Organization System), empower a more nuanced, network-centric approach to deciphering and systematising knowledge. These technologies enable the creation of intricate, interlinked knowledge repositories that reflect the complexity of human cognition, facilitating more sophisticated and context-sensitive processing of language by both humans and AI systems.

The purpose of this paper is to unravel the confluence of terminology management, semantic technologies and AI, spotlighting how this synergy can amplify the efficacy, precision and interoperability of translation and knowledge dissemination. We will scrutinise the evolution of concepts within terminology management, the role of semantic technologies in surmounting the challenges posed by the digital era and the potential of these technologies to bridge the rift between human and machine comprehension of language. Through this exploration, our goal is to furnish insights and pragmatic recommendations for practitioners in the field, contributing to the evolving discourse on the future trajectory of terminology management in an increasingly AI-influenced world.

2 The evolution of concepts and terminology management

The field of terminology management has undergone a remarkable evolution, particularly in the ways concepts are defined and used. This transformation is not just a shift in definition but represents a paradigmatic change in our understanding and handling of knowledge.

2.1 Conceptual shift: From static thoughts to dynamic knowledge

The traditional terminology management approach, notably ISO 1087:1990¹, focused on building a collection of isolated, static concept boxes. Each concept was defined and understood in isolation, with little regard for its relationship with other concepts in the field. This approach resulted in a fragmented understanding of knowledge, where concepts lacked the interconnectedness and dynamic nature that they possess in reality.

The advent of ISO 1087:2019² marked a significant shift in terminology management approach. This updated framework embraced the dynamic nature of knowledge by redefining concepts as ‘units of knowledge created by a unique combination of characteristics’. This nuanced shift transformed concepts from isolated markers into interconnected nodes within a vast network of understanding.

This knowledge-centric perspective highlighted the importance of concept systems, which are the interconnected networks of concepts that form the backbone of any subject matter. Unlike the 1990 framework, ISO 1087:2019 explicitly addresses concept systems, recognising their role in shaping our understanding of the world and the importance of managing them effectively.

While the older framework focused on compiling and organising concepts as distinct elements, the newer one emphasises the creation and management of concept systems. This shift in focus reflects the recognition that concepts are not static entities but dynamic nodes in a complex network of knowledge.

By embracing concept systems, terminology management practices can move beyond mere term cataloguing and contribute to a deeper understanding of the dynamic nature of knowledge. This, in turn, can lead to more effective communication, improved decision-making, and a more holistic approach to knowledge sharing. This approach mirrors the advancements in semantic technologies, where the emphasis is on the relationships and interdependencies between concepts, enabling more precise and comprehensive knowledge representation.

2.2 The pivotal role of semantic technologies

Semantic technologies have become the key to this evolved approach to terminology management. They enable the creation of structured, interconnected concept networks, transcending traditional linear or hierarchical models. Key tools and standards such as RDF, OWL and SKOS are central to this transformation, allowing for the depiction of concepts within a dynamic, interlinked web, thus enriching the scope and depth of knowledge management.

¹ See <https://www.iso.org/standard/5591.html>

² See <https://www.iso.org/standard/62330.html>

. Moreover, innovative developments such as Ontolex-Lemon³, a semantic framework designed to represent lexical information in relation to ontologies⁴, have emerged. This model bridges the gap between lexical and conceptual data, enabling seamless integration of terminologies with semantic web technologies. This ensures that linguistic data is not only machine-readable but semantically enriched, facilitating a deeper understanding and management of terminologies. This alignment of lexical resources with ontological concepts represents a significant advancement, paving the way for more nuanced, context-aware applications in AI and natural language processing.

2.3 Embracing a dynamic, interconnected knowledge perspective

The move from isolated terms to a networked concept framework aligns with broader trends in AI and human cognition. It signifies a paradigm shift from viewing knowledge as a collection of discrete data points to understanding it as a complex network of insights and information that can also be inferred. This shift is more than technical; it represents a philosophical reorientation towards a nuanced, context-aware and comprehensive approach to knowledge and language.

In essence, the evolution in concept definition and management symbolises a broader transformation in the field of terminology management. Driven by advancements in semantic technologies, this shift marks a transition from a static, isolated approach to knowledge to a dynamic, interconnected one. It highlights the need to recognise the relational and contextual dimensions of concepts, fostering a more effective and comprehensive approach to knowledge management in the digital era.

3 Semantic technologies: A needed change of paradigm

In the digital era, semantic technologies have emerged as a revolutionary force, fundamentally redefining the way we handle and interpret language. This shift transcends mere technological advancement; it represents a profound paradigm shift in the essence of linguistic processing and knowledge management. These innovative technologies are designed to comprehend and process the semantics of language, akin to human cognitive abilities, a stark contrast to traditional data processing methods that are limited to rigid, isolated data structures. Semantic technologies, instead, embrace the nuances of context and the intricacies of relationships, allowing for a more dynamic and intuitive interpretation of data.

Semantic technologies represent a switch in our understanding and manipulation of language. Moving beyond the traditional focus on syntax, these technologies delve into the deeper meanings and context of language, evolving from a two-dimensional perspective to a multi-dimensional one, where context, culture and intent are seamlessly integrated with terms and phrases.

Key technologies driving the semantic web, or Web 3.0, are well defined and include:

³ <https://www.w3.org/2019/09/lexicog/>

⁴ “ontology is an explicit specification of a conceptualization”, see <https://tomgruber.org/writing/ontolingua-kaj-1993.pdf>

The Resource Description Framework (RDF): A model for data interchange on the Web, extending the web's linking structure and forming a foundation for diverse knowledge representation. Data structure *meaning* is expressed by RDF that builds on XML principles for data interchange. This simple standard modelling language describes the distributed data on the Web in a syntax independent way allowing a meaningful indexing.

An RDF-based model expresses the meaning of its structure in the form of triples where things in the world are encoded like in a typical syntax analysis (subject, predicate and object) but by means of URIs that provide a machine-readable description.

The Web Ontology Language (OWL): Derived from RDF, OWL enables the explicit definition and interrelation of terms in vocabularies, essential for web ontology content.

The Simple Knowledge Organisation System (SKOS): A framework for structuring concept schemes like thesauri, classification schemes and taxonomies, enhancing controlled vocabularies⁵ with rich semantic relationships.

Semantic technologies confer numerous advantages:

Enhanced data interoperability, offering a universal framework for data sharing and reuse beyond the confines of individual applications, enterprises and communities,

Improved data linking capabilities, enabling the connection of diverse data sets in contextually meaningful ways,

Advanced data search and retrieval, leveraging the meaning and context of terms for more nuanced and accurate search functionalities,

Facilitation of knowledge discovery, extracting valuable insights from vast, unstructured datasets, revealing patterns and relationships not immediately discernible.

Embraces the Open World Assumption: Semantic technologies, such as OWL (Web Ontology Language), operate under the paradigm of the Open World Assumption. This assumption implies that information on the Web remains true until proven otherwise, and that the available information at one specific moment cannot be considered exhaustive. It is always to be assumed that new information may be discovered, potentially rendering previous information on the Web obsolete.

In terminology management, these technologies are crucial for:

Conceptual interconnectivity: They transition terminological data from isolated entities to a rich network of interconnected concepts, reflecting a more comprehensive understanding of language and knowledge,

Cognitive and linguistic alignment: These technologies enable the organisation of terminological data in a way that resonates with human cognitive processes and linguistic patterns, multilingual and cross-cultural communication: They play a pivotal role in navigating the complexities of multilingual terminology, ensuring the preservation of subtle semantic nuances across various languages.

⁵ See <https://link.springer.com/article/10.1057/dam.2010.29>

4 Overcoming challenges

In the dynamic landscape of language and communication, the digital era opens the way for a transformative shift. For linguists and language professionals, this evolution signifies not merely a change in tools but a profound reimagining of our approach to terms and meaning.

4.1 Revolutionising terminology management in the digital landscape

The digital revolution in terminology management surpasses the constraints of traditional lexicons. Advanced computational linguistics now unlock unparalleled levels of accessibility and precision. Terminologists in particular gain real-time access to a vast, continually evolving repository of languages, nuanced with cultural and contextual awareness.

However, traditional terminology management has long faced challenges, especially in areas like interoperability, data sharing and organising knowledge effectively. Legacy systems, with their compartmentalised approach, often struggled to dynamically interact with the ever-evolving nuances of linguistic and domain-specific contexts.

4.2 Facilitating semantic interoperability: Bridging linguistic divides

A key challenge in traditional management is semantic interoperability – ensuring diverse systems and organisations can meaningfully exchange and use data. Advances in semantic technologies, using standardised models like RDF and OWL, are bridging these gaps through their open and expandable approach. They enable systems to communicate in a universal (formal) language, facilitating seamless data integration and exchange across different platform and tools.

Digital terminology platforms facilitate unparalleled collaboration among language professionals. This interconnectedness ensures a consistency in language use, critical for maintaining the integrity and clarity of communication across various languages and dialects.

4.3 Data sharing: Breaking down silos of knowledge

Traditional systems frequently battled fragmented knowledge bases. Semantic technologies, with their networked architecture, are instead creating interconnected knowledge ecosystems. They ensure that updates or additions to the network are shared system-wide, giving all users access to the latest, comprehensive data.

The organisation of extensive terminological data remains a challenge, but semantic technologies are helping here by introducing ontologies as a structured framework for knowledge organisation. These ontologies do more than categorise terms; they reveal the intricate relationships between them, deepening our understanding of terminology.

In addressing these traditional challenges, Knowledge Organization Systems (KOS) are invaluable. They ease the creation, maintenance and use of semantically enriched terminological resources. KOS support robust multilingualism, ensuring terms are translated with context in mind, ensuring they adapt meaningfully rather than in a straightforward one-to-one manner.

4.4 Streamlining terminology management: Automation and optimisation

Semantic technologies bring a transformative efficiency to terminology management, automating and optimising various processes. They provide capabilities like automatic

synonym suggestion, change tracking and cross-platform term updates. Such automation not only saves time but also minimises human error, leading to more reliable and efficient management.

The digital dimension also allows for adaptive learning algorithms that cater for the specific needs of each professional. Whether it is a translator needing contextual nuances or an interpreter requiring rapid access to specialised terms, the system learns and adapts, offering personalised assistance.

4.4 Implications for language professionals

For language professionals, this shift encompasses both technical and philosophical dimensions. It offers linguists and language professionals new tools for capturing nuances and context. Semantic technologies enhance precision and add layers that can serve to accommodate cultural sensitivity, providing a nuanced understanding of idiomatic expressions, colloquialisms and cultural references, crucial in a globally connected yet culturally diverse world.

The synergy between human expertise and semantic technologies promises enhanced capabilities, combining the art of language with the precision of technology for richer, more accurate and culturally attuned communication.

While promising transformative opportunities, semantic technologies also present challenges such as the need for standardised methodologies, managing data complexity and scalability, and safeguarding data privacy and security. Data quality becomes even more relevant for terminologists due to the impact their work has on the much wider environment. Ongoing research and development are essential to address these challenges and maximise the potential of semantic technologies in terminology management and related fields.

4.5 A case in point: The European Union's semantic leap

For our community, this digital shift is not just about adopting new tools; it is about rethinking our roles in a world where language is increasingly fluid and interconnected. We must embrace these changes, leveraging the power of digital terminology to enhance our work's accuracy, efficiency and impact.

A practical illustration of overcoming these traditional challenges is evident in the workflow⁶ implemented by the Publications Office of the European Union and the Directorate-General Eurostat of the European Commission, in collaboration with the IATE Management Group and the Terminology Coordination Unit of the Council of the European Union. This workflow, focused on the maintenance of the corporate Countries and territories data asset⁷, leverages semantic technologies for enhanced knowledge management. Stakeholders, and re-users in particular, have observed marked improvements in data

⁶ See [Practical implementation: IATE and VocBench](#)

⁷ See <https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/country>

consistency and overall efficiency – a testament to the real-world benefits of semantic technologies.

On the other hand, the introduction and refinement of a virtual assistant named Publio⁸ on the Publications Office web portal (OP Portal⁹) represents a significant advancement. Publio is an Artificial Intelligence (AI)-powered tool designed to interact seamlessly with users, conducting searches based on keywords sourced from publicly available information on the OP Portal. Using a sophisticated language model, Publio comprehends user inquiries and aids in navigating their search process. This language model, integral to Publio's functionality, is continually enhanced and trained through user feedback. This iterative process ensures that Publio increasingly aligns with user expectations and effectively handles the variety of inputs in the three supported languages.

A noteworthy aspect of Publio's architecture is its integration of controlled vocabularies, specifically EuroVoc¹⁰, the multilingual thesaurus covering the activities of the European Union and various reference data assets, like authority tables¹¹, which are standardised reference list of codes and labels used to harmonise and facilitate data exchange across different systems and institutions. These components are diligently maintained and published by the Publications Office in collaboration with all EU institutions. This integration underscores the commitment to providing accurate, relevant, and comprehensive search experiences for users accessing the OP Portal.

5 Bridging semantic technologies and AI

In this pivotal section, we explore the symbiotic relationship between semantic technologies and artificial intelligence (AI), a duo that is redefining the landscape of language professions. This transformation is based on the interaction between meaning and information technology, which should revolutionise the way linguists, language professionals and knowledge managers approach their work in an increasingly digital world.

In the rapidly evolving landscape of digital linguistics¹², the confluence of semantic technologies and artificial intelligence marks a pivotal era. This synergy not only enhances both domains but also heralds a transformative phase in understanding and processing natural language. Semantic technologies, which imbue data with meaningful structure, are indispensable for AI's proficiency in language-related tasks. This collaboration is paramount in propelling advancements in Natural Language Processing (NLP) and Neural Machine Translation (NMT), two critical facets of AI that are central to decoding and generating natural language.

⁸ See <https://op.europa.eu/en/web/webtools/publio-the-publications-office-virtual-assistant> and <https://op.europa.eu/en/web/about-us/explainability-notice>

⁹ See <https://op.europa.eu/en/home>

¹⁰ See <https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc>

¹¹ See <https://op.europa.eu/en/web/eu-vocabularies/authority-tables>

¹² See <https://www.degruyter.com/serial/dil-b/html>

5.1 Semantic enrichment: The linguistic scaffolding for AI

Semantic enrichment, the process of augmenting digital content with metadata, is a cornerstone in this alliance. This enrichment transcends mere data categorisation, offering AI systems the contextual and nuanced understanding essential for accurate language interpretation. It also provides the formalisation of curated data that can make AI a trusted actor. For professionals in linguistics and translation, this means AI can provide translations with heightened precision, categorise content with greater relevance and retrieve information more efficiently. The nuanced understanding of context, idiomatic expressions and linguistic subtleties by AI, powered by semantically rich data, represents a significant leap forward in machine-assisted language and knowledge management services.

The integration of the nuanced understanding of semantics with the great computational ability of artificial intelligence creates a dynamic synergy. This convergence enables machines to not only process but also comprehend and contextualise language in ways previously unimaginable. For professionals in our fields, this means an unprecedented level of support in tasks ranging from context-sensitive translation to real-time interpretation.

5.2 Knowledge graphs and ontologies: Constructing AI's cognitive framework

At the heart of semantic technologies lie knowledge graphs and ontologies, instrumental in augmenting AI's cognitive capabilities. Knowledge graphs present a networked representation of data points, offering AI a structured knowledge base. Ontologies, in defining inter-concept relationships within specific domains, provide a blueprint for AI to decipher complex data relationships and contexts. This structured understanding is critical for AI's enhanced reasoning and decision-making, particularly in complex linguistic scenarios.

When harmonised with semantic technologies, AI enhances linguistic precision and diversity. We discuss how this integration facilitates a deeper understanding of idiomatic expressions, colloquial nuances and cultural contexts, thereby enriching our professional toolkit and enabling us to deliver more accurate and culturally sensitive translations.

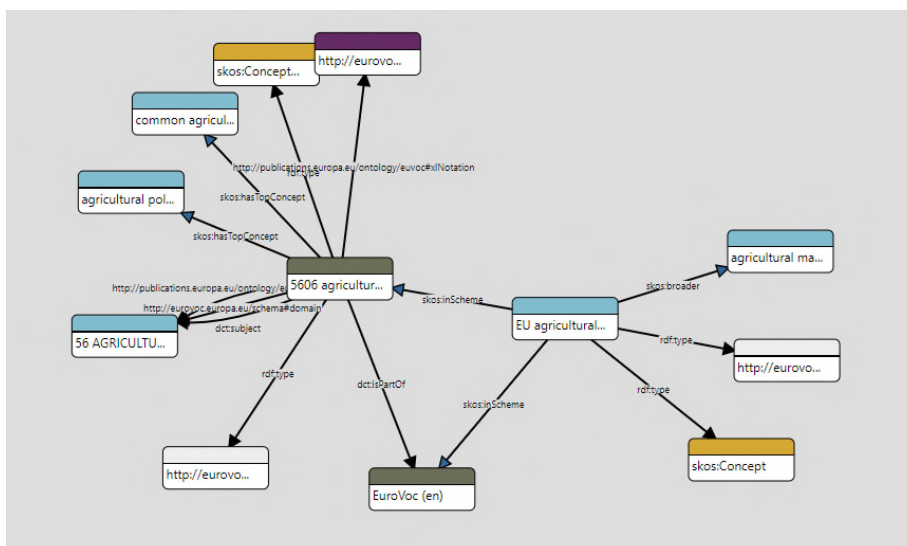


Figure 17. EuroVoc - Graph view of EU agricultural market

5.3 The essential function of terminologists

The role of terminologists is indispensable in the transition from traditional methods to AI-assisted approaches. This shift not only streamlines workflows but also paves the way for new possibilities in language sciences research and development. The crucial contribution of terminologists is accentuated, showcasing how their expertise acts as a cornerstone in enabling machines to better comprehend concepts.

In this integration, terminologists emerge as unsung architects. Their meticulous efforts in defining, managing, and categorising precise terminologies become the foundation of AI's learning process, especially in specialised domains. It is the terminologists' knowledge that ensures AI systems are equipped with accurately structured datasets, a critical element for the effective training and learning of AI algorithms. This collaboration exemplifies a unique intersection, where linguistic precision plays a pivotal role in advancing technology.

5.4 Practical applications and prospects

In practical applications, the fusion of semantic technologies and AI is already demonstrating remarkable outcomes. In information retrieval, AI, when powered by semantic technologies, delivers results that are not only relevant but also contextually nuanced. In neural machine translation, we can imagine that semantically enriched datasets would markedly improve translation accuracy, particularly for languages with intricate grammatical nuances.

Looking forward, the symbiosis between semantic technologies and AI is set for continuous evolution. As AI systems grow more sophisticated, their dependence on well-structured, semantically rich data will intensify. This interdependence will, in turn, fuel further advancements in semantic technologies, fostering a reciprocal growth cycle. This evolution will continually redefine the limits of AI in understanding and processing natural language, offering unprecedented opportunities for linguistics and language professionals to leverage AI in their work.

We anticipate that while AI will transform certain aspects of our roles, it will simultaneously amplify our ability to focus on the creative and nuanced aspects of language that remain uniquely human.

6 Practical implementation: IATE and VocBench

In the evolving landscape of digital linguistics, the collaborative integration of IATE¹³ (Interactive Terminology for Europe) and VocBench¹⁴ for the maintenance of the Countries and territories¹⁵ data asset represents a transformative leap in terminology management and semantic technology. IATE, a cornerstone in the European Union's linguistic framework, serves as a comprehensive multilingual terminology database. For linguists and translators, it offers an unparalleled repository of terms, serving as a vital reference point for accurate translations. Interpreters benefit from its extensive lexical database, enhancing their ability to

¹³ See <https://iate.europa.eu/>

¹⁴ See <https://op.europa.eu/en/web/eu-vocabularies/vocbench>

¹⁵ See <https://op.europa.eu/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/country>

convey nuanced meanings across languages. For terminologists, IATE provides a systematic framework for terminology management, fostering standardisation and uniformity across diverse linguistic landscapes. Its integration in a multidisciplinary workflow with VocBench, a cutting-edge, collaborative platform for managing semantic vocabularies, signals a revolutionary stride in our field.

VocBench introduces a new dimension in collaborative terminology management. It empowers communities of language experts to collectively curate and refine terminologies. This collaborative approach is pivotal for linguists, language professionals and knowledge managers, who strive to keep pace with evolving linguistic nuances and colloquialisms. VocBench's intuitive interface and real-time collaboration features ensure that terminologies remain current, relevant and reflective of contemporary linguistic usage.

This integration marks a paradigm shift, blending the robust, traditional database of IATE with the dynamic, semantic-rich capabilities of VocBench. Our objective has been to forge a workflow that not only safeguards the terminological data in IATE but also augments it with semantic intricacies and a web of interconnected knowledge. The resultant synergy is a testament to the power of combining established practices with innovative semantic technologies.

Central to this merger is the metamorphosis of a part of IATE's data into a structured, semantically interconnected knowledge base. VocBench's role here is pivotal; it transforms mere cataloguing into a process of creating meaningful, multidimensional semantic links. This new framework is invaluable for enhancing data interoperability and intuitive information retrieval, key aspects in our increasingly interconnected world.

Collaboration lies at the heart of VocBench, enabling a multitude of users to refine and enrich terminological data collectively. This joint approach not only ensures the continual evolution and accuracy of our terminologies but also embodies the democratic spirit of open-source development. Here, the voices of a broader community of experts resonate, enriching our linguistic tapestry.

Moreover, this integration addresses the crucial need for multilingual support, reflecting the European Union's diverse linguistic spectrum. VocBench's adequacy for managing multilingual vocabularies and ensuring contextually appropriate translations is unparalleled. The semantic enrichment tools it offers deepen our understanding of linguistic relationships, elevating the quality of terminological data to new heights.

The integration's most far-reaching impact, perhaps, is on AI and machine learning. The structured, semantically enriched data from IATE and VocBench significantly enhances the language comprehension capabilities of these technologies. This advancement is pivotal for improving machine translation, natural language processing and various AI-driven linguistic applications.

For linguists, language professionals and knowledge managers, this integration streamlines terminology management and fosters more coherent, efficient cross-cultural communication. As we look ahead, the continual evolution of this integration promises to keep in stride with the rapid advancements in semantic technology and AI, reshaping the landscape of terminology management and language processing in profound ways.

7 Conclusion: Embracing semantic technologies in terminology management for the AI era

As we conclude our exploration into the integration of semantic technologies within terminology management, several key insights and future directions emerge. This convergence not only redefines existing practices but also sets a new standard for the role of terminology in the digital era.

The rise of artificial intelligence and machine learning technologies has catapulted us into a new paradigm where traditional methods of terminology management, though foundational, are eclipsed by the exigencies of rapidity, precision and contextual nuance. Semantic technologies have emerged as the backbone in this transformative era, offering a depth of conceptual understanding and interconnectivity that mirrors the intricacies of human cognition. This shift from static terminology repositories to dynamic, interlinked knowledge networks is not just an incremental improvement; it is an indispensable evolution to remain abreast of the swift progress in AI.

A pivotal consequence of this integration is the democratisation of knowledge. By rendering complex terminologies accessible and comprehensible to both artificial and human intellects, we are effectively dismantling barriers to knowledge acquisition. This is most important in our globalised environment, where information must transcend linguistic and cultural divides to achieve universal comprehensibility.

Looking ahead, the trajectory of terminology management is unequivocally collaborative. It envisages a harmonious interplay among terminologists, thematic experts, knowledge managers, AI developers and semantic technology specialists. Terminologists are entrusted with the critical task of encapsulating the subtle nuances and contextual intricacies of terms. AI developers harness these insights to forge systems that are more astute and context aware. Semantic technologies provide the indispensable infrastructure for this collaborative venture, enabling a seamless fusion of these disparate fields.

As we navigate this new digital terrain, our traditional challenges are being reshaped into opportunities for growth, innovation and greater connectivity within our diverse linguistic communities. In embracing these advancements, we are not just keeping pace with change; we are leading it, ensuring that our work continues to be relevant, impactful and integral in the digital era.

In sum, the integration of semantic technologies in terminology management signifies a fundamental shift, effectively addressing and surmounting the hurdles of traditional approaches. By augmenting interoperability, simplifying data sharing and refining knowledge organisation, these technologies chart a course towards a more efficient, coherent and adaptable future in terminology management.

Nonetheless, this integration is not devoid of challenges. AI and semantic technologies are in a state of perpetual flux, necessitating continuous adaptation and learning. Ensuring consistency and standardisation across diverse languages and terminologies remains a formidable task. Yet, these hurdles also pave the way for ground-breaking innovation and development within the field.

In essence, the integration of semantic technologies in terminology management transcends a mere technical enhancement; it marks a pivotal stride towards actualising the full potential of AI in deciphering and processing the nuances of natural language. As we forge ahead, it is vital for professionals in this domain to not only adapt to these changes but to actively contribute to

the forging of a more interconnected and intelligible digital landscape. We are transitioning from simply capturing thoughts to encapsulating knowledge within terminologies – a journey full of promises for the future of information exchange and communication in the AI era.

References

- Fišer, Darja, Andreas Witt, and Clarin Eric. 2022. The Infrastructure for Language Resources
- Hedden, Heather. 2010. Taxonomies and controlled vocabularies best practices for metadata. *J Digit Asset Manag* **6**, 279–284 (2010). <https://doi.org/10.1057/dam.2010.29>
- Gruber, Thomas R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**(2):199-220, 1993.
- Dunaevsky, Carolina. 2015. Terminology and the Semantic Web - Paradigms, challenges and problems. (Master's thesis). Technical University of Cologne, Germany.

o

The GAMETRAPP Project: Post-editing Neural Machine Translation of Research Abstracts in a Gamified Environment

Cristina Toledo-Báez

IUITLM, University of Malaga

toledo@uma.es

Laura Noriega-Santiáñez

IUITLM, University of Malaga

laura.noriega@uma.es

Abstract

Given the continuous refinement of neural machine translation (NMT) systems, post-editing (PE) is increasingly present in the translation world. Furthermore, new educational realities have led to revolutionary learning techniques, such as gamification, which are already being tested in translation teaching. Against this background, the GAMETRAPP project is framed within the multilingual context and the need for scholars to disseminate science in English irrespective of their disciplines and L1 background. The GAMETRAPP project is funded by the Spanish Ministry for Science and Innovation (TED2021-129789B-I00) and its main goal is to bring the NMT + full PE of research abstracts closer to non-professional translators and scholars using a gamified environment. The project setup is based on the Iberian Spanish>American English language direction. After defining the NMT, PE, and gamification concepts, this article presents the methodology of the project, especially the research abstract collection and abstract processing. In addition, the design of the future gamified environment is also briefly explained. Finally, the conclusions reached in this first year of the project are detailed, as well as the future steps.

1 Introduction

The world of translation has experienced a tremendous change with the emergence of machine translation (MT) systems, and specially, neural machine translation (NMT) systems, which have reshaped multiple professional realities in different fields. Their arrival has developed new ways of conceiving translation practice and has led to the birth of post-editing (PE) to reach the appropriate quality standards for publishing an MT output (Vieira, 2019). PE is now part of the translator's workflow (Zaretskaya et al., 2016) and it allegedly saves time and improves quality (Herbig et al., 2019) when used in the L2>L1 combination (i.e., from second language into first language).

Directionality has received little attention in PE (Stasimioti et al., 2021) and research on the use of MT + PE has predominantly focused on L2>L1 direction. However, the directionality axiom is being questioned and different authors have explored PE into L2 by translators/post-editors, non-professional translators (Yamada, 2019) and scholars (Parra Escartín et al., 2017; O'Brien *et al.*, 2018; Parra Escartín and Goulet, 2020). Given the multilingual context and the need for scientific dissemination, scholars, regardless of their L1 backgrounds, are forced to publish their results in English, the *lingua franca* in academic writing (O'Brien et al., 2018). Based on this need, the GAMETRAPP project was born out of the idea that the use of MT+ PE into English as L2 is expected to improve academic output in scholars having Spanish as L1.

Another element of paramount importance in the GAMETRAPP project is the use of gamification, an increasingly popular learning technique that helps and motivates students to learn by means of playful activities (video games, escape rooms, treasure hunts, among others). Gamification is an increasingly used methodology that can be defined as “the use of game design elements in non-game contexts” (Deterding et al., 2011: 9), such as in professional settings, and it has proved to encourage students to learn and engage in the classroom content (Alsawaier, 2018). Indeed, gamified activities boost learning and memorization, as they

increase the level of satisfaction and decrease the students' level of stress (Gutiérrez-Artacho and Olvera-Lobo, 2016). Recent studies have already successfully applied gamification techniques with Translation and Interpreting students (cf. Gutiérrez-Artacho and Olvera-Lobo, 2016; Alcaide-Martínez and Taillefer, 2022), but the particularity of our project lies in two key points: (1) bringing NMT + full PE into L2 closer to scholars who are non-professional translators (2) by means of an ad-hoc gamified environment.

In the following paragraphs, the GAMETRAPP project is described in depth. Section 2 details the aims and the methodology of the project, followed by Section 3, focused on the collection of research abstracts. Section 4 elaborates on data processing combining human translation (HT), MT and PE whereas Section 5 briefly describes the gamified environment. Finally, Section 6 presents the main conclusions and future steps.

2 The GAMETRAPP project: Aims and methodology

The GAMETRAPP project brings together three very different yet complementary fields: NMT, full PE, and gamification. The main goal of the GAMETRAPP project, funded by the Spanish Ministry for Science and Innovation (TED2021-129789B-I00), is to bring the NMT + full PE closer to scholars from different fields using a gamified environment. Specifically, the main contribution is to create a responsive application (for web, mobiles, and tablets) for training on how to fully post-edit research abstracts. The potential user of this application is a scholar and/or non-professional translator having Spanish as L1 and an advanced competence in English as L2. The Iberian Spanish>American English language direction is used for the project setup. These two variants have been chosen because the Universities involved in the project are from Spain (University of Malaga, University of Cordoba, University Pablo de Olavide, University of Alcalá, Complutense University of Madrid, University of Valladolid, and Valencia International University) and United States (Kent State University and Utah Valley University).

The methodology for this project encompasses 3 main phases that are subdivided into 8 subphases and 30 tasks. The 3 main phases are: 1) Pre-use of gamified environment, 2) Use of gamified environment, and 3) Post-use of gamified environment. The project is currently in Phase 1, having some tasks already finished (Tasks 5, 6, 10, and 13) and some tasks in progress (Tasks 12 and 14). Tasks related to data collection (Task 5) are described in Section 3, whereas tasks focused on data processing (Tasks 6, 10, and 13) are explained in Section 4. Task 14 is detailed in Section 5. Previous tasks (Tasks 1-4) are focused on theoretical concepts relevant to both NMT and PE literacy, and gamification, but they are beyond the scope of this paper.

3 Data collection: Research abstracts

This data collection is based on Task 5, which dealt with the manual selection of source texts. Specifically, research abstracts presenting an IMRaD structure were collected. This acronym stands for Introduction, Methods, Results, and Discussion and represents a summary of the different sections of research articles (Fraser, 2002). Since the 20th century, the IMRaD structure has been used to write scientific abstracts to the point that it has become standardized in certain fields (Sollaci and Pereira, 2004). In fact, “editors and scientists agree that IMRaD

provides a consistent framework that guides the author to address several questions essential to understanding a scientific study” (Wu, 2010: 1346).

Regarding Task 5, having real and published research abstracts was of paramount importance. Therefore, Scimago Journal & Country Rank¹ (SJR) was chosen as the abstract database. Two criteria were used to filter journals: Open access journals 1), indexed in the first and second quartile in 2022, and 2) from Spain. A total of 244 journals met these criteria. Then, 5 research abstracts per journal were chosen, having a total of 1220 abstracts that were manually collected and classified in a Google Drive folder. All these abstracts were classified using the following 12 labels within an Excel sheet:

1) Record code (by subject area, sub-area, and abstract number), 2) Journal, 3) Quartile (Q1 or Q2), 4) Area (Humanities, Science, Social Sciences or Engineering), 5) Subject Area, 6) Category (both items used the subject areas and categories proposed by SJR), 7) Article reference (DOI), 8) URL (link to the article), 9) Type of abstract (“Structured” if it followed the IMRaD structure or “Non-structured” if this rhetorical structure is not followed), 10) Protocolized abstract (“Yes”, if it is previously labeled as “Structured”; “No” or “Partially”, if it is labeled as “Non-structured”, depending on whether the abstract follows the IMRaD structure partially or not at all), 11) Partially protocolized (if it is previously labeled as “Partially”, this label indicates the IMRaD section(s) the abstract contains), and 12) Number of words. Once the labeling task was finished, only abstracts meeting the following three selection criteria were considered: 1) abstracts having all the sections in the IMRaD structure; 2) published in 2023 and 3) written by authors affiliated with Spanish universities. The final result was 126 abstracts: 22 from Humanities, 17 from Sciences, and 87 from Social Sciences.

4 Data processing: HT, MT, and PE

Once the research abstracts were selected, Tasks 6, 10, 12, and 13 were performed. Task 6 was devoted to the selection of the NMT system. Google Translate was the NMT system chosen for this project, since it is an open access free online platform and several studies point it out as the most used NMT system amongst scholars (cf. Parra Escartín et al., 2017; O’Brien et al., 2018; Parra Escartín and Goulet, 2020).

Task 10 entailed having the research abstracts translated by a professional translator with English as L1. The resulting translations were used as a gold standard for both the PE and the design of gamified activities. Task 13 addressed the PE of research abstracts by a professional post-editor. A post-editor with English as L1 fully post-edited the 126 research abstracts machine-translated by Google Translate.

As to Task 12, the results from the three tasks, i.e., HT, NMT and PE outcomes, were analyzed by a group of GAMETRAPP researchers using an evaluation rubric, classifying the NMT errors and PE output based on the MQM (Multidimensional Quality Metrics)² errors typology and the Post-Edit Me! Project³, respectively. Then, these will be used to elaborate

¹Available at <https://www.scimagojr.com/>

² Available online at <https://themqm.org/error-types-2/typology/>

³ Available online at <https://oer.uclouvain.be/jspui/handle/20.500.12279/829>

patterns for full PE of research abstracts. The gamified activities will be designed on the basis of these patterns, as described in the following section.

5 Design of the gamified environment

Task 14 is devoted to the design of the gamified environment. As explained above, the gamified environment is based on the collection of HT, NMT, and PE data. The errors and difficulties encountered in the NMT output will be the input for the creation of the gamified activities. Specifically, the MQM errors, i.e. Terminology, Accuracy, Linguistic conventions, Style, Locale conventions, Audience appropriateness, Design and markup. In addition, the main linguistic, stylistic, and rhetorical features of research abstracts will be also considered when designing the activities.

Thanks to these activities, users will learn three main notions for full PE research abstracts: 1) pinpointing the NMT errors, 2) PE them by using the most suitable strategy and, then 3) identifying the PE errors. Thus, the gamified environment will be set to engage and train scholars on how to correct NMT errors based on real examples.

To create this educational scenario, a specialized gamification software will be used, namely Articulate, with the help of a hired professional expert in this field. Thus, an escape room experience is under construction based on the different parts of the IMRaD structure and NMT+PE literacy of research abstracts.

6 Conclusions

This paper presents the primary steps derived from the first year of the GAMETRAPP project, as well as the current and future implementation tasks. The completed tasks, both data collection and data processing, lay the foundations of the project as the 126 abstracts represent a real linguistic sample. Indeed, NMT and PE taxonomies help to analyze the data output and, thus create full PE notions of research abstracts following the IMRaD structure. Hence, this planned methodology sets the design of tailor-made activities based on the most current needs of non-professional translators and researchers.

Concerning the pending tasks, the project is working on the creation of a gamified environment that integrates activities to help users recognize and solve the different types of NMT errors as well as the main linguistic and rhetorical features of research abstracts. Thus, users will be trained on some notions to fully PE research abstracts from Iberian Spanish into American English through engaging activities. Therefore, the gamified environment within the application is expected to help scholars feel more confident when post-editing potential publications in their L2. In addition, this app can also be used during the translators' learning process and as a way of raising awareness of the difficulty of PE of a text without expertise.

Once Phase 1 is finished, we will then continue with Phase 2, which entails the use of the gamified environment. In this phase, non-professional translators, specifically scholars, will test the application. Finally, Phase 3 will focus on the post-use of the gamified environment by means of testing and user surveys.

Acknowledgements

The GAMETRAPP project (ref. no. TED2021-129789B-I00/AEI/10.13039/501100011033/ Unión Europea NextGenerationEU/PRTR) is funded by the Spanish Ministry for Science and

Innovation under the Ecological Transition and Digital Transition Call 2021. This research was also carried out in the framework of the research projects: NEUROTRAD (B1-2020_07), VIP II (PID2020-112818GB-100/AEI/10.13039/501100011033), T2T (D5-2023_14), RECOVER (ProyExcel_00540), DIFARMA (HUM106-G-FEDER, 2024-2025) and TRADUTEACH (PIE22-14).

References

- Alcaide-Martínez, Marta, and Lidia Taillefer. 2022. Gamification for English language teaching: A case study in translation and interpreting. *Lebende Sprachen*, 67(2): 283–310.
- Alsawaier, Raed. 2018. The effect of gamification on motivation and engagement. *International Journal of Information and Learning Technology*, 35(1): 56–79.
- Deterding, Sebastian, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: Defining 'gamification'. In *Proceedings of the 15th International Academic MindTrek Conference. Envisioning Future Media Environments*, pages 9–15.
- Fraser, Simon. 2002. A Statistical Analysis of the Vocabulary of Medical Research Articles (2): Differences across the "IMRAD" Structure. *看護学統合研究*, 4(1): 27–34.
- Gutiérrez-Artacho, Juncal, and María Dolores Olvera-Lobo. 2016. Gamification in the Translation and Interpreting Degree: A New Methodological Perspective in the Classroom. In *EDULEARN16 Proceedings*, pages 50–58.
- Herbig, Nico, Santanu Pal, Josef Van Genabith, and Antonio Krüger. 2019. Multi-Modal Approaches for Post-Editing Machine Translation. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, USA*, pages 1–11.
- O'Brien, Sharon, Michel Simard, and Marie-Josée Goulet. 2018. Machine Translation and Self-Post-Editing for Academic Writing Support: Quality Explorations. In Joss Moorkens, Sheila Castilho, Federico Gspari, and Stephen Doherty, editors, *Translation Quality Assessment. Machine Translation: Technologies and Applications*. Springer, pages 237–262.
- Parra Escartín, Carla, and Marie-Josée Goulet. 2020. When the Post-Editor is not a Translator: Can machine translation be post-edited by academics to prepare their publications in English? In Maarit Koponen, Brian Mossop, Isabelle S. Robert, Giovanna Scocchera, editors, *Translation Revision and Post-Editing*. Routledge, pages 89–106.
- Parra Escartín, Carla, Sharon O'Brien, Marie-Josée Goulet, and Michel Simard. (2017). Machine Translation as an Academic Writing Aid for Medical Practitioners. In *Proceedings of MT Summit XVI*, pages 254–267.
- Sollaci Luciana. B., and Mauricio G. Pereira. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J Med Libr Assoc*, 92: 364–367.
- Stasimioti, Maria, Vilelmini Sosoni, and Konstantinos Chatzitheodorou. 2021. Investigating post-editing effort. Does directionality play a role?. *Cognitive Linguistic Studies*, 8 (2): 378–403.
- Vieira, Lucas Nunes. 2019. Post-editing of Machine Translation. In Minako O'Hagan, editor, *The Routledge Handbook of Translation and Technology*, Routledge, pages 319–336.

- Wu, Jianguo. 2011. Improving the writing of research papers: IMRAD and beyond. *Landscape Ecol* 26: 1345–1349.
- Yamada, Masaru. 2019. Language learners and non-professional translators as users. In Minako O'Hagan, editor, *The Routledge Handbook of Translation and Technology*, Routledge, pages 183–199.
- Zaretskaya, Anna, Mihaela Vela, Gloria Corpas Pastor, Miriam Seghiri. 2016. Comparing Post-Editing Difficulty of Different Machine Translation Errors in Spanish and German Translations from English. *International Journal of Language and Linguistics*, 3(3): 91-100.

Google Translate Error Analysis for Mental Healthcare Information: Evaluating Accuracy, Comprehensibility, and Implications for Multilingual Healthcare Communication

Jaleh Delfani

University of Surrey, UK
j.delfani@surrey.ac.uk

Constantin Orăsan

University of Surrey, UK
c.orasan@surrey.ac.uk

Hadeel Saadany

University of Surrey, UK
hadeel.saadany@surrey.ac.uk

Özlem Temizöz

University of Surrey, UK
o.temizoz@surrey.ac.uk

Eleanor Taylor-Stilgoe

University of Surrey, UK
e.j.taylor-stilgoe@surrey.ac.uk

Diptesh Kanojia

University of Surrey, UK
d.kanojia@surrey.ac.uk

Sabine Braun

University of Surrey, UK
s.braun@surrey.ac.uk

Barbara Schouten

University of Amsterdam, Netherlands
b.c.schouten@uva.nl

Abstract

This study explores the use of Google Translate (GT) for translating mental healthcare (MHealth) information and evaluates its accuracy, comprehensibility, and implications for multilingual healthcare communication, through analysing GT output in the MHealth domain from English to Persian, Arabic, Turkish, Romanian, and Spanish. Two datasets comprising MHealth information from the UK National Health Service website and information leaflets from The Royal College of Psychiatrists were used. Native speakers of the target languages manually assessed the GT translations, focusing on medical terminology accuracy, comprehensibility, and critical syntactic/semantic errors. GT output analysis revealed challenges in accurately translating medical terminology, particularly in Arabic, Romanian, and Persian. Fluency issues were prevalent across various languages, affecting comprehension mainly in Arabic and Spanish. Critical errors arose in specific contexts, such as bullet-point formatting, specifically in Persian, Turkish, and Romanian. Although improvements are seen in longer-text translations, there remains a need to enhance accuracy in medical and mental health terminology and fluency, whilst also addressing formatting issues for a more seamless user experience. The findings highlight the need to use customised translation engines for MHealth translation and the challenges when relying solely on machine-translated medical content, emphasising the crucial role of human reviewers in multilingual healthcare communication.

1 Introduction

The World Health Organisation (2019)¹ reported that around 970 million people worldwide, or 1 in 8 individuals, faced mental disorders, primarily anxiety and depression. The COVID-19 pandemic exacerbated these statistics, revealing a significant 26% increase in anxiety disorders and a notable 28% rise in major depressive disorders within a year (WHO, 2022). In the same vein, the UK Mental Health Foundation² highlighted that untreated mental health issues contribute to 13% of the global disease burden. According to their projections, by 2030, mental health problems, especially depression, are expected to become the leading cause of both mortality and morbidity globally.

Despite effective prevention and treatment options, the majority of those with mental disorders lack access to adequate care, especially migrants and refugees who may not speak the language of the country they are trying to settle in (Krystallidou *et al.*, 2024). In healthcare settings, there are challenges to accessing human interpreters such as waiting times, financial constraints, and limited availability (Al Shamsi *et al.*, 2020). In other instances, where information is available in written format with translation, instead of interpretation, there is a need for automated translation which is occasionally employed (Turner *et al.*, 2019; Chen and Acosta, 2016; Taylor-Stilgoe *et al.*, 2023).

Machine translation (MT) has emerged as a potentially valuable tool to overcome language barriers in healthcare, offering access to vital information for individuals with limited language proficiency. Generic MT tools like GT provide free access to automatic translation across many languages, but these translations vary in quality, thus raising concerns about reliability, liability, and data privacy, especially in safety-critical situations (Vieira *et al.*, 2021).

This paper explores the errors introduced by GT when used to access mental health-related materials such as website information and digital leaflets. It is organised as follows: Section 2 offers a brief literature review on the use of technology to facilitate communication in healthcare settings, particularly when participants lack a common language or when users need to comprehend a document written in an unfamiliar language. Section 3 outlines the methodology adopted to select data, translate it into other languages, and conduct an error analysis. Section 4 presents the findings related to two scenarios where GT was employed for data translation. The paper concludes with final remarks and suggestions for future research.

2 Literature Review

Effective communication in mental healthcare, especially in multilingual situations, is extremely important. Language serves as a channel for understanding, empathy, and successful treatment. Moreover, effective multilingual communication dismantles cultural barriers, minimises stigma, and cultivates a sense of inclusivity. Studies indicate that migrants and refugees face an increased risk of developing depression and anxiety disorders due to exposure to stressors following resettlement, limited social support, and societal stigma and discrimination (Rousseau and Frounfelker, 2019). Furthermore, research provides evidence indicating higher prevalence rates of specific mental health disorders (*e.g.*, posttraumatic stress

¹ <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>

² <https://www.mentalhealth.org.uk/>

and psychosis-related disorders) among migrant populations compared to non-migrant populations (Priebe *et al.*, 2016). Language barriers can impede effective communication regarding treatment requirements and available care choices among patients, resulting in reduced utilisation of psychiatric healthcare services (Doğan *et al.*, 2019; Kiselev *et al.*, 2020; Krystallidou *et al.*, 2024; Marquine and Jimenez, 2020; Ohtani *et al.*, 2015). The presence of stigma and a hesitancy to seek assistance further exacerbates the language barrier (Giacco *et al.*, 2014). Overcoming language barriers is essential for achieving high levels of satisfaction among both medical professionals and patients, ensuring proper treatment, and maintaining patient safety (Al Shamsi *et al.*, 2020). There are two primary solutions to address language barriers: using interpreting services and leveraging available translation applications.

Interpreting services may escalate both the cost and duration of the treatment process (Al Shamsi *et al.*, 2020) and might not be consistently accessible (Arafat, 2016; Doğan *et al.*, 2019; Felsman *et al.*, 2019; Khanom *et al.*, 2021; Pallaveshi *et al.*, 2017; Shrestha-Ranjit *et al.*, 2017). In such scenarios, machine translation emerges as the most readily available solution.

Generic tools like GT or bespoke solutions (Dew *et al.*, 2018; Haddow *et al.*, 2021; Vieira *et al.*, 2021) play a crucial role in easing communication between healthcare professionals and patients who lack a shared language. In such scenarios, direct interaction occurs between the patient and healthcare provider by means of a device and interface granting access to machine translation software. This dynamic interaction takes place in co-located settings where patients and providers use smartphones or tablets to access translation software. Alternatively, it may occur in situations where they are connected remotely through communication technology, as exemplified by telehealth consultations employing integrated translation tools such as Skype Translator³.

Apart from the insufficient exploration of the complex interactions involved in the use of machine translation tools in interpersonal healthcare communication, a significant obstacle to their practical implementation is the issue of quality. Currently, machine translation fails to provide accurate mediation for numerous language pairs in diverse healthcare settings. To tackle the existing challenges associated with MT, while still incorporating a degree of automation, various semi-automated approaches, have been developed in particular phrase-based translation apps such as Xprompt and BabelDr (Braun *et al.*, 2023). These apps are typically pre-loaded with validated human translations of common phrases and sentences, providing essential communication support in specific healthcare settings. The interaction with these apps can be as intricate as the interaction with pure MT software.

Over the past decade, publicly available generic machine translation tools have shown improvement. Translation applications like GT and Microsoft Translator now provide the translation of written and spoken input into text and/or speech output in near-real time for an expanding range of language pairs. The use of generic MT tools in daily clinical practice became apparent in a study investigating attitudes toward vaccination among Polish and Romanian communities in England. A significant number of healthcare professionals delivering vaccines to these communities reported relying on free MT tools to communicate with people who did not speak English (Moberly, 2018a). Although official guidance in the UK does not endorse their use in medical consultations, healthcare workers perceived these tools as more accessible than professional interpreting services, especially in time-pressured appointments. In response to this discovery, medical advisers emphasised the potential risks of

³ <https://www.skype.com/en/features/skype-translator/>

using tools like GT in everyday clinical practice, citing the possibility of introducing communication errors and compromising patient safety, which could expose doctors to legal action (Moberly, 2018b). However, the advisers acknowledged that MT tools might have a limited role in emergencies or other exceptional circumstances.

GT stands out as one of the most recognised and extensively used machine translation tools among the general public. Supporting translations for 133 languages (as of May 2022)⁴ and compatible with both iOS and Android systems, this free application is a popular choice. Platforms like X (formerly Twitter) often rely on GT to offer translation services, and users are well-acquainted with its functionality. According to reports from 2021, the tool translates over 100 billion words daily,⁵ indicating that the public is inclined to use it for their translation needs. Research indicates that refugees frequently employ GT on their smartphones as their primary online translation tool (Abujarour, 2022) and it serves as the most accessible and free primary means of communication in healthcare settings where language is a barrier. Nevertheless, it is crucial to grasp the accuracy and potential drawbacks of GT output, especially when dealing with sensitive and critical healthcare content (Leite *et al.*, 2016).

In evaluating the effectiveness of GT in translating emergency department discharge instructions from English to Spanish and Chinese, Khoong *et al.* (2019) discovered that a substantial percentage of sentences were accurately translated (92% for Spanish and 81% for Chinese). However, they noted that 2% of Spanish and 8% of Chinese sentence translations revealed the potential for significant or life-threatening harm, primarily due to errors in word disambiguation. In a parallel study examining additional language pairs, Taira *et al.* (2021) observed that GT output was inconsistent across six language pairs, with accuracy rates ranging from 55% to 94%. Assessing another generic translation app, iTranslate, in translating common questions posed by diabetes patients to clinicians, Chen, Acosta, and Barry (2017) found that the MT output was comparable to human translation in terms of accuracy for simple sentences but error-prone for complex sentences.

The National Health Service (NHS) in England explicitly advises its staff against using online MT services, citing the lack of quality assurance regarding the translations (NHS England, 2023). Nevertheless, instances are prevalent where healthcare staff resort to non-specialised, commercially available MT tools, such as GT, when providing interpersonal or written assistance to patients with limited to no proficiency in the English language (Bell *et al.*, 2020; Moberly, 2018a; Royal College of Midwives, 2017). Vieira *et al.*, (2021) highlight that research on the implications of the widespread, and potentially uninformed, use of this technology remains limited. Studies investigating the impact of MT on patient medical record documentation reveal that healthcare professionals are largely unaware of the errors that MT can introduce, particularly concerning abbreviations (Taylor-Stilgoe *et al.*, 2023).

Considering these concerns and knowledge gaps, the objective of this investigation is to explore the inaccuracies introduced by GT when translating materials related to mental health from English to five languages, each with differing levels of resources. By scrutinising the accuracy and potential pitfalls of MT output in this critical healthcare context, this research

⁴ <https://blog.google/products/translate/24-new-languages/>

⁵ <https://ttcwetranslate.com/how-does-google-translate-work/>

seeks to contribute insights into the nuanced challenges and opportunities associated with the use of MT tools in mental health communication.

3 Methodology

This section outlines the methodology employed in this paper for the preparation of the investigated datasets, the utilisation of the machine translation engine, and the assessment of translation quality.

3.1 Datasets

For our study, we used two distinct datasets: one comprising isolated sentences and the other full documents, both written in English and containing information related to mental health. This approach was chosen to explore the potential impact of context, allowing a clearer comparison of results between sentence-level and document-level analysis.

The first set of sentences was extracted from the UK NHS website⁶, which provides healthcare information to patients. This website was chosen for its comprehensive resources on health conditions, symptoms, and treatments. It features a guide, crafted by healthcare professionals, that offers insights into a variety of health issues, advising visitors on what actions to take and when to seek assistance. From these articles, we extracted 100 English sentences (1494 words) related to the mental health domain, which were then translated into other languages using GT. We will refer to this dataset as the “NHS dataset”.

The second dataset was constructed using digital information leaflets sourced from the UK Royal College of Psychiatrists⁷. We will refer to this dataset as the “RCP dataset”. These leaflets are originally written in English and present user-friendly and evidence-based information on mental health problems, treatments, and related subjects. Qualified psychiatrists, with input from patients and carers, contribute to the creation of these informative materials. For our experiments, we selected the leaflet with the topic of “Depression” (1267 words). We used GT to translate our datasets into five languages under investigation in this study, namely Persian, Modern Standard Arabic and Turkish (low-resourced languages), Romanian (a medium-resourced language), and Spanish (a high-resourced language). Our objective was to evaluate GT’s performance in the mental health context across languages with varying levels of resources in two scenarios: a) translating individual sentences and b) translating longer texts (contextualised paragraphs).

3.2 Data Preparation

The NHS dataset was translated into the aforementioned languages using GT and organised into separate spreadsheets. Subsequently, native speakers of each respective language conducted manual analyses (the analysis procedure is described in the next section). For the second dataset, which comprises digital leaflets, we opted to provide context to GT to assess whether its performance differed from the first scenario where individual sentences were translated. To be more specific, in our study, GT was stress-tested on two types of data: a)

⁶ <https://www.nhs.uk/conditions/>

⁷ <https://www.rcpsych.ac.uk/mental-health>

individual sentences which may lack the overall context; b) longer stretches of text from leaflets as contextually coherent units.

Throughout all the experiments detailed in this paper, we employed the online version of GT, without any customisation or tuning for a specific domain. The translations were made using the version available in June 2023.

3.3 Evaluation Method

Assessing the output of machine translation is a challenging task that has undergone extensive examination. Commonly employed evaluation methods include automatic approaches such as BLEU (Papineni *et al.*, 2002) and METEOR (Banerjee and Lavie, 2005), which automatically compare machine-generated output with a reference translation. While these methods prove valuable in the development of machine translation systems, they fail to elucidate why a translation falls short. Consequently, we opted for manual error analysis. Despite its time-consuming nature, this approach is invaluable in discerning the specific errors made by MT engines. Following a preliminary analysis of the MT output for our datasets, we developed an error taxonomy to systematically capture the most significant errors in the mental health context. In designing our error annotation scheme, we drew inspiration from existing typologies such as MQM. However, we focused specifically on (critical) errors that can affect the mental health message, rather than addressing the full spectrum of errors covered in MQM.

Our error typology comprised the following:

- **Inaccuracy of mental health and medical terminologies:**

Instances where either or both of these aspects were not translated accurately, and therefore, could have consequences for the effectiveness of the message. To ensure consistency and reliability in identifying such terms, the sentences in our datasets were reviewed by cross-referencing them with several resources, including the World Health Organisation's key terms and definitions in mental health,⁸ NHS mental health conditions⁹, and Bupa mental health glossary¹⁰. It should be mentioned that 53 of the 100 sentences in our NHS dataset included such terms.

- **Syntactic/semantic errors:**

These errors may arise from incorrectly constructed target sentences or inaccuracies in translating words or phrases.

- **Comprehensibility issues:**

Translations that are intricate and challenging for individuals with diverse levels of mental health literacy to understand.

- **Fluency issues:**

These refer to problems in the translation that disrupt the natural flow and ease of reading. A translation with fluency issues might be grammatically correct but still feel awkward or unnatural to a native speaker, making the content less readable or smooth.

⁸ <https://www.who.int/southeastasia/health-topics/mental-health/key-terms-and-definitions-in-mental-health>

⁹ <https://www.nhs.uk/mental-health/conditions/>

¹⁰ <https://www.bupa.co.uk/~media/files/mms/bins-02812.pdf>

- **Clarity and Coherence issues:**

These occur when the translation does not convey the intended meaning clearly and logically, leading to confusion or ambiguity. Clarity and coherence issues can make the text difficult to follow, causing a lack of logical flow that connects ideas seamlessly. This may affect the reader's ability to fully understand the message.

- **Culturally insensitive translations:**

Renderings that are deemed unacceptable due to the cultural nuances of the target language. This type of error is especially crucial, considering the divergent perspectives on mental health prevalent across various cultures.

- **Critical errors:**

These are errors that significantly impact the meaning of the message and may pose a risk to patient health outcomes.

For our error analysis, we recruited native speakers of the target languages to review the MT output alongside the source texts in English. One evaluator was assigned for each language, and these evaluators shared similar profiles: all had a background in linguistics, extensive experience in machine translation research, and fluency in English. The same evaluators assessed both datasets, ensuring consistency across the evaluations. The evaluators were responsible for identifying translation errors and categorising them into predefined classes. They were thoroughly briefed on the error classification guidelines to ensure consistency in their assessments. While it is acknowledged that employing multiple evaluators per language and ensuring identical profiles across evaluators would enhance the reliability of the analysis, this approach was not feasible within the scope of this study nevertheless, this limitation was considered when interpreting the results. It should also be noted that the analysis presented in this paper focused on the linguistic aspects of the translation and the readers' comprehension of the message, rather than on evaluating the quality of the translation from a healthcare perspective. Such an evaluation is slated for future consideration.

4 Findings

The results will be presented in two parts: the first part will discuss the findings collectively, i.e., based on the phenomena we observed across languages, while the second part will showcase more detailed findings with examples related to each language under investigation. It should be noted that different types of errors were counted separately to ensure that each was recognised individually. To maintain consistency and comparability across and within languages, clear and detailed guidelines were provided to the evaluators. These guidelines outlined specific criteria and included examples and clarifications for identifying and categorising errors, ensuring a uniform approach across all evaluations.

4.1 Collective Results for the NHS Dataset

After analysing the GT output for 100 sentences from the NHS website, it was observed that Arabic exhibited the highest error rate in almost all categories among the other languages under investigation (Figure 1). This was followed by Persian and Romanian, respectively. The quality of the Spanish translation was not as good as that of Turkish. Turkish had the lowest number of errors among the languages we investigated in this dataset.

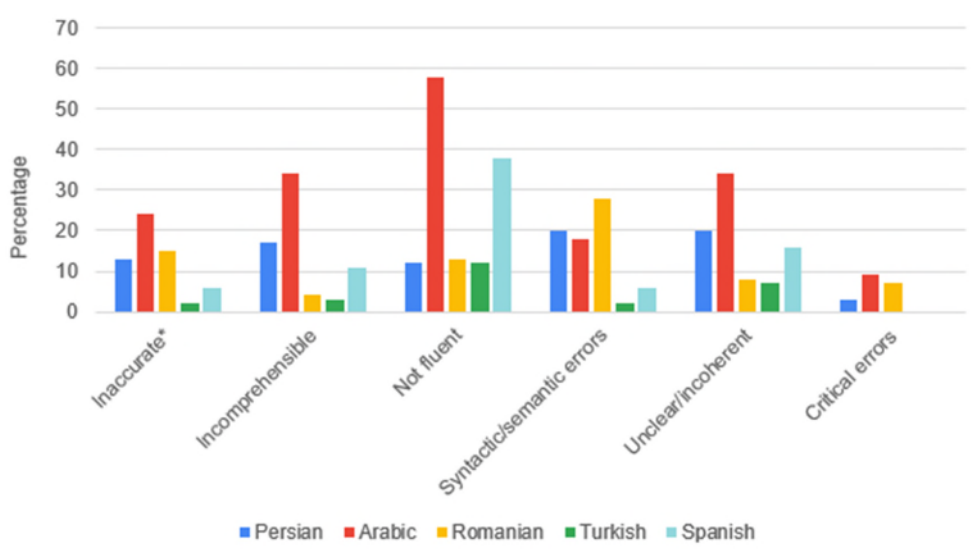


Figure 1. Relative frequency of the error types in all languages- NHS dataset

As seen in Table 1 below, Arabic (shown in red) exhibited the highest number of errors in each category, except for the syntactic/semantic errors, where Romanian had the highest error rate. Seven of these errors in Romanian could pose a risk to the patient’s life. Critical errors were also observed in Arabic and Persian, with Arabic having the highest number of this error type. Fluency issues were among the most frequently occurring errors in almost all languages, even in Spanish, which is considered a high-resourced language. Turkish demonstrated the lowest error rate in almost all categories among other languages (shown in green), which is rather surprising, as this language falls into the same category of low-resourced languages as Arabic and Persian. It is important to consider the possibility that the lower error rate could, in part, be influenced by the evaluator’s specific approach or interpretation of the error categories, even though the evaluator was a native speaker of Turkish. While we have no specific evidence to suggest that this was the case, it is worth acknowledging this potential factor in interpreting the results. Further analysis with multiple evaluators could help to rule out any evaluator-related biases and confirm the robustness of these findings. Notably, no cases of cultural insensitivity were observed in GT output for this dataset.

Error Type	Persian	Arabic	Romanian	Turkish	Spanish
Inaccurate terminology	13	24	15	2	6
Incomprehensible	17	34	4	3	11
Not fluent	12	58	13	12	38
Syntactic/semantic errors	20	18	28	2	6
Incoherent	20	34	8	7	16
Culturally insensitive	0	0	0	0	0
Critical errors	3	9	7	0	0

Table 1: Absolute frequency of the error types in all languages- NHS dataset

4.2 Collective Results for the RCP Dataset

For this particular dataset, we opted for a contextual analysis, focusing on paragraphs rather than individual sentences, a method distinct from our approach with the other dataset. This involved using a Word document to systematically examine each paragraph, annotating errors observed. Given the shift to a paragraph-level analysis, we initially considered the introduction of new error types to account for issues such as inconsistencies across sentences and incorrect coreferences. However, we ultimately decided to adapt our existing error taxonomy, emphasising the identification of these issues within the broader context of our established categories. The evaluators were provided with clear directives to ensure consistent application of the error taxonomy across the dataset, despite the complexity of paragraph-level analysis. They documented the prevalent error types in each language and, where possible, identified the potential causes of these errors. This approach allowed us to maintain coherence in our analysis while addressing the specific challenges posed by paragraph-level content.

The findings demonstrated a notable divergence from those obtained with the NHS dataset. Arabic exhibited a superior translation quality, with only minor pronoun errors detected. In contrast, Persian presented challenges in punctuation, bullet point formatting, and code-switching in the GT output, leading to syntactic/semantic errors, incomprehensibility, and coherence issues. Both Romanian and Turkish exhibited issues stemming from the bullet point structure, contributing to syntactic and semantic errors, comprehensibility problems, and fluency issues. The decline in the quality of GT output for Turkish, when translating paragraphs compared to individual sentences in the other dataset, was unexpected. This decline was mainly caused by the bullet-point structure of the source text i.e., the English information leaflets. Spanish, in this dataset, revealed problems related to comprehensibility, fluency, coherence, lack of gender agreement, and incorrect/missing abbreviations.

4.3 Outcomes Specific to Each Language

4.3.1 Persian

NHS dataset

In the analysis of the NHS data, 53 sentences contained medical or mental health terminology. Persian translations showed 13% inaccuracies in such terms, with 3 cases rendering messages incomprehensible and 2 affecting overall fluency. Of 100 Persian translations, 17 were incomprehensible, 12 had fluency issues, and 20 exhibited syntactic/semantic errors, 3 of which were critical. In addition, 20 incoherent sentences were documented.



Figure 2. GT error analysis for Persian

The following are illustrative examples of the above-mentioned errors identified within this dataset.

Error type	Source text	GT output	Back translation
Comprehensibility	Give a fractional pause after each expiration and inspiration .	پس از هر انقضا و الهام، مکث کسری بدهید.	After each expiry and inspiration , give a fractional pause.
Critical	The diet of persons suffering from depression should completely exclude tea, coffee, alcohol, chocolate and cola, all white flour products, sugar, food colourings, chemical additives, white rice and strong condiments.	رژیم غذایی افراد مبتلا به افسردگی باید به طور کامل شامل چای، قهوه، الکل، شکلات و کولا، تمام محصولات آرد سفید، شکر، رنگ های غذایی، افزودنی های شیمیایی، برنج سفید و چاشنی های قوی باشد.	The diet of persons suffering from depression should completely include tea, coffee, alcohol, chocolate and cola, all white flour products, sugar, food colourings, chemical additives, white rice and strong condiments.

Table 2. Examples of errors for GT output in Persian- NHS dataset

In the first example provided, the terms ‘expiration and inspiration’ are translated as ‘the expiry date’ and ‘being inspired to do something’, rather than ‘exhale and inhale’ in this context. This interpretation can certainly result in a misunderstanding of the original message, impacting its comprehensibility and clarity. In the second example above, falling into the category of critical error cases, the term ‘exclude’ has been translated to its complete opposite, ‘include’, which can pose a risk to patient safety and health outcome.

RCP dataset

The main challenges faced by GT in translating Persian mental health-related leaflets were related to two key issues: problems with punctuation, particularly in translating bullet point formatting, and the occurrence of code-switching between Persian and Latin scripts in the translated content. These challenges primarily led to syntactic errors, resulting in a significant loss of comprehensibility and clarity in the translated text, as well as compromised linguistic fluency. The analysis identified six critical errors in the dataset, highlighting the need for improvement in handling these specific issues. Table 3 below illustrates one such error.

Error type	Source text	GT output	Back translation
Critical	In your mind, you lose your self-confidence, start to feel hopeless, and perhaps even suicidal.	در ذهن شما، شما اعتماد به نفس خود را از دست بده شروع به احساس ناامیدی و شاید حتی خودکشی کن.	In your mind, lose your confidence, start to feel hopeless, and even kill yourself.

Table 3: Example of errors for GT output in Persian- RCP dataset

In the example mentioned above, a critical error is observed in the translation. The original sentence, which provides information regarding the symptoms of depression, has been translated in an imperative sense, urging the patient to lose their confidence, start feeling disappointed, and maybe even commit suicide. This is an extremely serious error that could potentially lead to the patient contemplating self-harm or suicide. Such critical errors pose a significant risk to the well-being and safety of the patients.

4.3.2 Arabic

NHS dataset

In the English to Arabic translation, 24% of the 53 medical terms analysed were mistranslated, a higher rate than in Persian. The lack of fluency in the Arabic dataset was notably high at 58%, and 34% of translated sentences were incomprehensible. Arabic translations exhibited a relatively higher number of critical errors, where syntactically correct sentences provided incorrect information in the target language. Examples include the translation of ‘mantras’ as singing, leading to a loss of essential mental health advice, and the reversal of advice from “practice yoga and meditation” to “avoid yoga and meditation,” creating challenges in detecting errors due to the fluency and syntactic correctness of the Arabic sentences (Table 4).

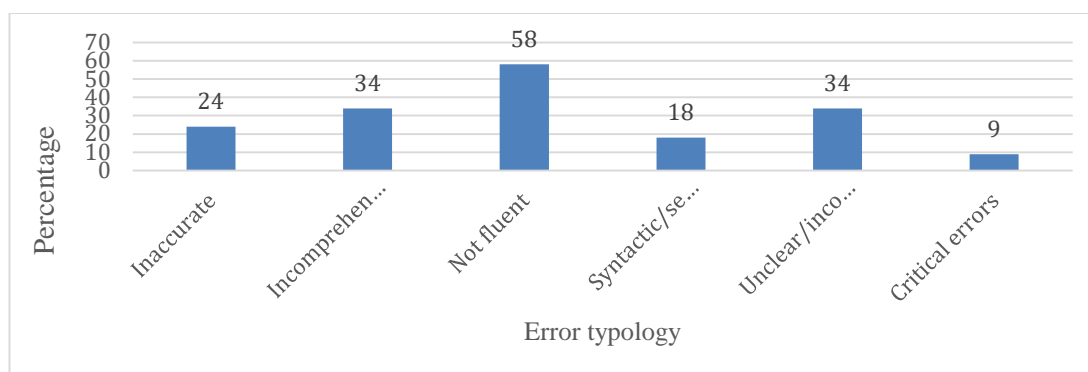


Figure 3: GT error analysis for Arabic

Error type	Source text	GT output	Back translation
Semantic	Focus your mind on mantras or breathe .	ركز عقلك على التغمي أو التنفس.	Focus your mind on singing and breathing
Critical	Practice yoga or meditation to avoid stress in life.	تجنب ممارسة اليوجا أو التأمل ضغط في الحياة.	Avoid yoga or meditation on stresses of life.

Table 4. Examples of errors for GT output in Arabic- NHS dataset

RCP dataset

Unlike the translation of NHS sentences, English to Arabic translation of longer medical leaflets showed a higher standard. The Arabic translation of the depression leaflet was fluent and comprehensible, with minor errors involving pronoun choices. Overall, the performance of GT was notably better with longer text spans in the translation from English to Arabic.

4.3.3 Turkish

NHS dataset

GT’s Turkish output was high-quality with no critical errors. The most common issues were related to fluency, accounting for 12% of the translation output. However, these fluency problems did not impact the overall clarity or comprehensibility of the content. Although, 7% of the translated sentences were deemed unclear or incoherent, and 3% were incomprehensible, only 2% of sentences containing mental health terminology were inaccurately translated and these were not considered critical.

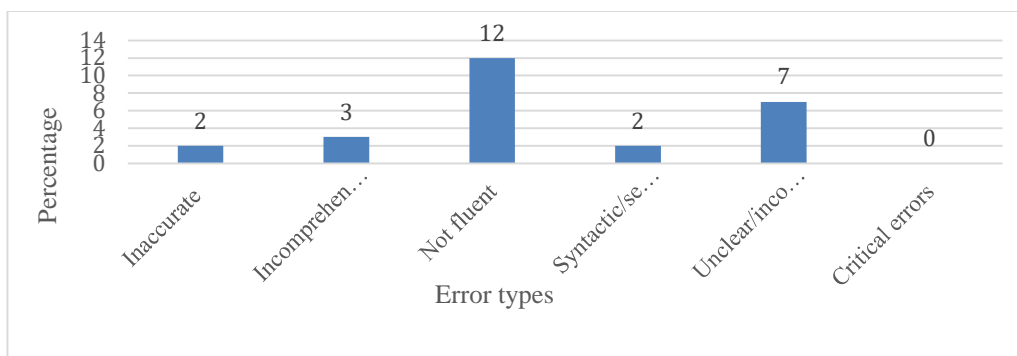


Figure 4. GT error analysis for Turkish

Error type	Source text	GT output	Back translation
Semantic	Resist the temptation to drown your sorrows with alcohol.	alkolle Acılarınızı cazibesine boğmanın direnin.	Resist the temptation to drown your sorrows with alcohol.

Table 5. Example of errors for GT output in Turkish- NHS dataset

In the above example, the infinitive ‘to drown’ is literally translated as ‘boğmanın’ which would be appropriate ‘to drown someone in water’. However, in the given context, a fluent translation would require the use of ‘Acılarınızı alkolle bastırmanın cazibesine direnin’ (resist the temptation to suppress your sorrows with alcohol).

RCP dataset

For this dataset, GT yielded lower quality in the Turkish translation as to the NHS dataset. While there are no critical errors, the bullet-point structure caused fluency issues that affected readability. This structure resulted in parts of the information regarding the symptoms of depression being translated as instructions rather than descriptions, which could lead to confusion. Medical and mental health-related terms were translated accurately to Turkish; however, inconsistency was observed related to the rendition of medical acronyms in this dataset. For example, the phrases such as ‘cognitive behavioural therapy (CBT)’ and ‘selective serotonin reuptake inhibitor (SSRI)’ are translated correctly as ‘bilişsel davranışçı terapi (CBT)’ and ‘seçici serotonin geri alım inhibitörü (SSRI)’, respectively; however, the acronyms are left untranslated. On the other hand, in one instance, the acronym is translated correctly when it is used without the expanded version, i.e., ‘CBT programmes’ as ‘BDT programları’. In another example, however, both the therapy name and the acronym were correctly translated; for example, Electroconvulsive therapy (ECT) is translated as Elektrokonvülsif tedavi (EKT).

4.3.4 Romanian

NHS dataset

The predominant error type in GT output for Romanian was syntactic/semantic errors, constituting 28 percent of the errors, with 7% classified as critical errors. For example, in Table 6 below, the word “high”, translated as “big” in Romanian, fails to convey that the source sentence discusses being high due to drugs. Out of a total of 53 sentences containing medical/mental health terminology, 7 of them (13%) misinterpreted one or more terms within the sentence. Nevertheless, in most cases, while the translations may seem peculiar, they are

likely to provide the reader with an understanding of the intended meaning, and therefore, they are not deemed critical errors.

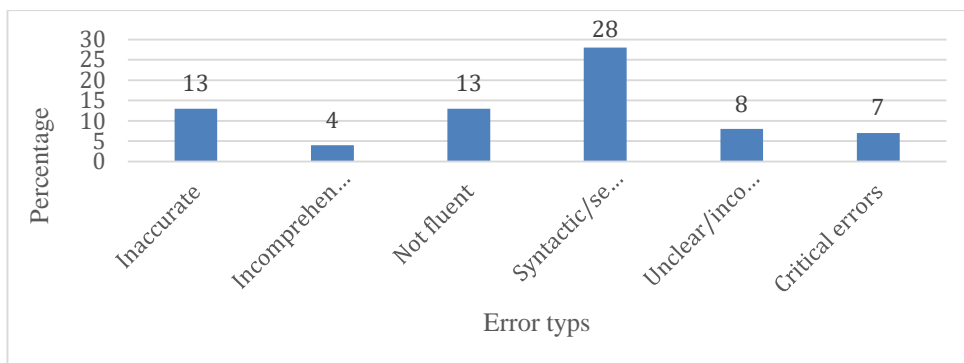


Figure 5. GT error analysis for Romanian

Error type	Source text	GT output	Back translation
Semantic	You may also have to encourage the person to visit the psychiatrist or ask for urgent help whenever the sufferer is so "high" that he or she is no longer aware that anything is wrong.	De asemenea, este posibil să fiți nevoit să încurajați persoana să viziteze psihiatrul sau să cereți ajutor urgent ori de câte ori suferința este atât de "mare" , încât nu mai este conștient de faptul că ceva nu este în regulă.	In addition, it is possible that you <u>have to</u> encourage the person to visit the psychiatrist or to ask for urgent help whenever the suffering is so "big" that he is no longer aware that something is not right.

Table 6. Examples of errors for GT output in Romanian-NHS dataset

RCP dataset

The analysis of the leaflets revealed numerous errors in GT Romanian output, mainly due to bullet point formatting. Errors included verb/pronoun disagreements, incorrect verb forms, and distorted sentences. The formatting issue led to critical errors, especially in translating from second person to third person or infinitive forms. For instance, the translation of “[you] can’t eat and lose weight” (as symptoms of depression) results in an inaccurate message: “they can’t eat and cannot lose weight”. Moreover, translation errors can occasionally cause a shift in focus from the reader to a broader audience, causing confusion. These issues are primarily linked to bullet point formatting, as paragraphs without special formatting were translated with fewer mistakes.

4.3.5 Spanish

NHS dataset

For the English to Spanish output, 3 of the 53 medical terminology instances were considered to have been rendered inaccurately, affecting 6% of the corresponding translations. In 2 of the 3 cases, the inaccuracy of the medical terminology negatively impacted the comprehensibility of the intended message, though none were considered as critical errors. While 38 out of 100 sentences were observed as containing instances of disfluency, only 6 contained semantic or

syntactic errors that could impact the target reader’s understanding of the mental health message.

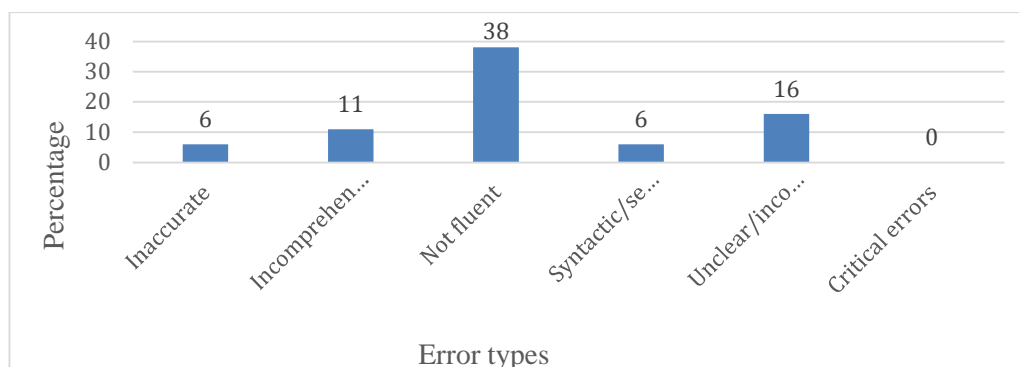


Figure 6. GT error analysis for Spanish

Error type	Source text	GT output	Back translation
Semantic	Sleep is disturbed when you have a maladjusted family .	El sueño se molesta cuando tienes una familia desadyectada .	Sleep is disturbed when you have a [incomprehensible] family .

Table 7. Example of errors for GT output in Spanish- NHS dataset

In the given passage, the adjective ‘desadyectada’ is highlighted as a significant impediment to the quality of translation. This term, used to render ‘maladjusted’, not only fails to convey the original meaning but also appears to be entirely invented by GT.

RCP dataset

The translation of medical leaflets from English to Spanish faced significant challenges, including inconsistencies in subject pronouns and verb conjugations. Issues included mismatches between informal and formal ‘you’ (syntactic errors), incorrect verb forms (syntactic errors), and inappropriate word choices (semantic errors). Gender agreement problems, incorrect abbreviations, untranslated terms, and missing articles were also noted. However, in the absence of bullet point structure in the source text, the text quality improved.

5 Concluding Remarks

Our research underscores the critical importance of recognising and addressing the limitations inherent in the use of Google Translate within the context of mental health. It is crucial to adopt a cautious approach and implement necessary precautions to safeguard patient well-being and facilitate effective communication. One prominent aspect that emerges from our findings is the pressing need for substantial enhancements in GT’s performance, particularly within the realm of mental health. This becomes even more imperative when considering languages with limited resources such as Arabic and Persian, where GT may exhibit shortcomings that could impede communication and understanding and potentially pose a risk to patient well-being and safety. Recognising and addressing these issues promptly and advocating for improvements in the accuracy of translations is essential, especially in areas that involve intricate structures, such as bullet points, code-switching, and specialised medical and mental health terminology.

Furthermore, it is crucial to acknowledge that GT is susceptible to errors at any given point, even within high-resourced languages, and should not be the sole resource, or relied upon exclusively, in sensitive contexts such as mental health. Recognising and embracing the collaborative role of human reviewers is integral to the responsible and effective use of machine translation tools like GT in the mental health context. However, while advocating for better translation performance and the implementation of human revision is essential, it is also important to consider the feasibility and challenges of these measures. In many current conditions, particularly in resource-constrained environments, the integration of human reviewers may be difficult due to factors such as limited availability of qualified translators, time constraints, and budget limitations. To successfully implement these safe-guards, there would need to be significant investment in training and recruiting skilled professionals, as well as the development of efficient workflows that allow for the timely review of translations.

5 Future work

Our research is ongoing and is part of a broader research initiative. Possible areas for improving this research include:

- Broadening the scope of our language coverage and augmenting the sample size. By encompassing a more extensive array of languages and increasing the number of evaluators in our study, we aim to obtain a more comprehensive understanding of the nuances involved.
- Undertaking a comparative analysis, for instance, comparing the performance of GT versus ChatGPT. This comparative approach will allow us to determine strengths, weaknesses, and potential areas for refinement in this context.
- Conducting case studies, involving real users in specific scenarios and practical applications for issues that may not be apparent through quantitative analysis alone, can be explored to gain further insights.
- Examining the output generated by GT within the mental healthcare domain, from the mental healthcare professionals' perspective, in the analysis process. This can be done to assess whether the errors identified have the potential to impact patient health outcomes, shifting the focus beyond linguistic elements alone.

Acknowledgments

This project was funded by the European Union Asylum, Migration and Integration Fund [Award No:101038491]. The views and opinions expressed in this paper are solely those of the authors and do not necessarily represent the official views of the funding agency. The funding agency is not responsible for the ideas expressed or conclusions drawn in this manuscript.

References

- Abujarour, Safa. 2022. Integration through education: Using ICT in education to promote the social inclusion of refugees in Germany. *Journal of Information Systems Education*, 33(1), pages 51-60.
- Al Shamsi, Hilal, Abdullah G. Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for healthcare: A systematic review. *Oman medical journal* 35, no. 2: e122.
- Arafat, Nahed Munir. 2016. Language, culture and mental health: A study exploring the role of the transcultural mental health worker in Sheffield, UK. *International Journal of Culture and Mental Health*, 9(1), pages 71–95.
- Banerjee, Satanjeev, and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65-72.
- Bell, Sadie, Vanessa Saliba, Gail Evans, Stephen Flanagan, Sam Ghebrehewet, Helen McAuslane, Bharat Sibal, and Sandra Mounier-Jack. 2020. Responding to measles outbreaks in underserved Roma and Romanian populations in England: the critical role of community understanding and engagement. *Epidemiology & Infection* 148, pages 1-8.
- Braun, Sabine, Khetam Al Sharou, and Özlem Temizöz. 2023. Technology use in language-discordant interpersonal healthcare communication. *The Routledge Handbook of Public Service Interpreting*, pages 89-105.
- Chen, Xuewei, Sandra Acosta, and Adam Etheridge Barry. 2016. Evaluating the accuracy of Google translate for diabetes education material. *JMIR diabetes*, 1(1), p.e5848. <https://diabetes.jmir.org/2016/1/e3/>
- Chen, Xuewei, Sandra Acosta, and Adam E. Barry. 2017. Machine or human? Evaluating the quality of a language translation mobile app for diabetes education material, *JMIR Diabetes*, 2(1).
- Dew, Kristin N., Anne M. Turner, Yong K. Choi, Alyssa Bosold, and Katrin Kirchhoff. 2018. Development of machine translation technology for assisting health communication: A systematic review, *Journal of Biomedical Informatics* 85 (September), pages 56–67.
- Doğan, Nareg, Gül Dikeç, and Ersin Uygun. 2019. Syrian refugees' experiences with mental health services in Turkey: "I felt lonely because I wasn't able to speak to anyone". *Perspectives in psychiatric care* 55(4), pages 673-680.
- Felsman, Irene C., Janice C. Humphreys, and Rebecca Kronk. 2019. Measuring distress levels of refugee women to enhance community-based psycho-social interventions. *Issues in Mental Health Nursing* 40(4), pages 310-316.
- Giacco, Domenico, Aleksandra Matanov, and Stefan Priebe. 2014. Providing mental healthcare to immigrants: Current challenges and new strategies. *Current opinion in psychiatry* 27(4), pages 282-288.
- Haddow, Barry, Alexandra Birch, and Kenneth Heafield. 2021. Machine translation in healthcare. In *The Routledge Handbook of Translation and Health*, Routledge, pages 108-129.

- Khanom, Ashrafunnesa, Wdad Alanazy, Lauren Couzens, Bridie Angela Evans, Lucy Fagan, Rebecca Fogarty, Ann John et al. 2021. 'Asylum seekers' and 'refugees' experiences of accessing health care: A qualitative study. *BJGP open* 5(6).
- Khoong, Elaine C., Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA internal medicine* 179(4), pages 580-582.
- Kiselev, Nikolai, Naser Morina, Matthis Schick et al. 2020. Barriers to access to outpatient mental health care for refugees and asylum seekers in Switzerland: The therapist's view. *BMC Psychiatry* 20(1). BioMed Central Ltd. <https://doi.org/10.1186/s12888-020-02783-x>
- Krystallidou, Demi, Özlem Temizöz, Fang Wang, Melanie de Looper, M., Emilio Di Maria, Nora Gattiglia, Stefano Giani, Graham Hieke, Wanda Morganti, Cecilia Serena Pace, Barbara Schouten, Sabine Braun. 2024. Communication in refugee and migrant mental healthcare: A systematic rapid review on the needs, barriers and strategies of seekers and providers of mental health services, *Health Policy*, <https://doi.org/10.1016/j.healthpol.2023.104949>
- Leite, Fernando Ochoa, Catarina Cochat, Henrique Salgado, Mariana Pinto da Costa, Marta Queirós, Olga Campos, and Paulo Carvalho. 2016. Using Google Translate in the hospital: A case report. *Technology and Health Care* 24(6), pages 965-968.
- Marquine, María J., and Daniel Jimenez. 2020. Cultural and linguistic proficiency in mental health care: A crucial aspect of professional competence. *International psychogeriatrics* 32(1), pages 1-3.
- Moberly, Tom. 2018a. Doctors choose Google Translate to communicate with patients because of easy access, *The British Medical Journal*, 362: k3974. <https://doi.org/10.1136/bmj.k3974>.
- Moberly, Tom. 2018b. Doctors are cautioned against using Google Translate in consultations, *The British Medical Journal* 363: k4546. <https://doi.org/10.1136/bmj.k4546>.
- Ohtani, Ai, Takefumi Suzuki, Hiroyoshi Takeuchi, and Hiroyuki Uchida. 2015. Language barriers and access to psychiatric care: a systematic review. *Psychiatric Services* 66(8), pages 798-805.
- Pallaveshi, Luljeta, Ahmed Jwely, Priya Subramanian, Mai Odelia Malik, Lueda Alia, and Abraham Rudnick, A. 2017. Immigration and Psychosis: An Exploratory Study. *Journal of International Migration and Integration*, 18(4), pages 1149–1166.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Priebe, Stefan, Domenico Giacco, and Rawda El-Nagib. 2016. Public health aspects of mental health among migrants and refugees: a review of the evidence on mental health care for refugees, asylum seekers and irregular migrants in the WHO European Region. Copenhagen: WHO Regional Office for Europe; (Health Evidence Network (HEN) Synthesis Report 47).
- Rousseau, Cécile, and Rochelle L. Frounfelker. 2019. Mental health needs and services for migrants: An overview for primary care providers. *Journal of Travel Medicine* 26(2).

- Royal College of Midwives. 2017. Stepping up to Public Health: A new maternity model for women and families, midwives and maternity support workers. Royal College of Midwives, London.
- Shrestha-Ranjit, Jagamaya, Elizabeth Patterson, Elizabeth Manias, Deborah Payne, and Jane Koziol-McLain. 2017. Effectiveness of primary health care services in addressing mental health needs of minority refugee population in New Zealand. *Issues in Mental Health Nursing* 38(4), pages 290-300.
- Taira, Breena R., Vanessa Kreger, Aristides Orue, and Lisa C. Diamond. 2021. A pragmatic assessment of google translate for emergency department instructions. *Journal of General Internal Medicine* 36(11), pages 3361-3365.
- Taylor-Stilgoe, Eleanor J., Constantin Orăsan, and Félix do Carmo. 2023. An Exploration of Risk in the Use of MT in Healthcare Settings with Abbreviations as a Use Case. *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*.
- Turner, Anne M., Yong K. Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L. Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the usefulness of translation technologies for emergency response communication: a scenario-based study. *JMIR public health and surveillance* 5(1). <https://doi.org/10.2196/11171>.
- Vieira, Lucas Nunes, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society* 24(11), pages 1515-1532.

Updating translator education programs: Adapting to technologies and their impacts in the Canadian language industry

Elizabeth Marshman

School of Translation and
Interpretation, University of
Ottawa/OLST

Elizabeth.Marshman@uOttawa.ca

Anwar Alfetlawi

School of Translation and
Interpretation, University of Ottawa

aalfe022@uOttawa.ca

Haifa Ben Naji

School of Translation and
Interpretation, University of Ottawa

hbenm070@uOttawa.ca

Dipen Dave

Telfer School of Management,
University of Ottawa

ddave@uOttawa.ca

Ahmed Elhuseiny Bedeir

School of Translation and
Interpretation, University of Ottawa

aelhu037@uottawa.ca

Ting Liu

School of Translation and
Interpretation, University of Ottawa

tliu109@uOttawa.ca

Abstract

Professionalizing translator education programs must strive to prepare graduates for a quickly evolving field in which technologies are constantly changing and, in turn, affecting workflows, tasks, and required competences. In doing so, they are subject to the challenges of updating curricula quickly enough, and of keeping up to date with the evolving perceptions and expectations of stakeholders, from prospective students to employers. This case study describes some of the data gathered during a 2023 market study in the context of program reform at the University of Ottawa's School of Translation and Interpretation. By exploring stakeholders' priorities, we hope to provide insights into curriculum design and recruitment that may be useful for other programs with similar goals.

1 Introduction

Translator education is challenging, given the need to facilitate students' acquisition of numerous competences during a compact program, and to prepare graduates for current and future industry needs. The slow evolution of university programs, compared to the more agile private sector, increases these challenges (Austermuehl, 2013; Sánchez-Castany, 2023). More than ever, we are experiencing rapid evolution and integration of technologies into the industry. In addition to computer-aided translation (CAT) tools already well-established in many programs (e.g., Austermuehl, 2013; Bowker & Marshman, 2010; European Master's in Translation Group, 2022; Rodríguez-Castro, 2018), machine translation (MT)—particularly neural machine translation (NMT)—is now also essential in training (e.g., Massey &

Ehrensberger-Dow, 2017; Mellinger, 2017). The advent of widely available generative AI (genAI) tools based on large language models (e.g., ChatGPT) has also suggested new applications in the workplace and in translator training (e.g., Pierce, 2023; Pym, 2023; Yamada, 2023a, 2023b), and provoked debate over their potential impact and reliability (Roy & Poirier, 2023; van der Meer, 2023).

In this paper, we will review findings of a market study carried out between May and October 2023 to guide program reform—specifically the creation of a professionalizing Master of Translation (MTr)—at the University of Ottawa’s School of Translation and Interpretation. We hope that this case study of translator education programs will offer insights relevant to recruiting students and preparing them for complex and evolving workplaces, in Canada and abroad.

We will discuss the following questions:

How are technology developments shaping stakeholder views of the language industry?

How important are technological competences in translator education programs?

What kinds of technological skills and knowledge are most important to target, and why?

When and where should technologies be introduced in translator education programs?

2 Background and context

2.1 Scholarship in translator education and technologies

Given the increasingly important role of technologies in the translation industry, their integration into translator education programs has elicited extensive attention. One major question surrounds *which* competences should be acquired, and how changes in the industry affect these choices. Clearly, students should be prepared to deal with both MT and CAT, but how much to focus on each is uncertain (e.g., Austermuehl, 2013; Sánchez-Castany, 2023). Moreover, the technology-related competences students are expected to acquire require careful consideration. Most agree that rather than focusing on the mechanical manipulation of tools, students should be prepared to think critically about technologies and how they affect translators and their working processes (e.g., Sánchez-Castany, 2023; Vandaele, 2017), and about important ethical questions (Bowker, 2020; Massey & Ehrensberger-Dow, 2017; Moniz & Parra Escartín, 2023; Moorkens, 2022). Moreover, they must develop capacities recognized as beyond the capacity of (current) technologies, including creativity (e.g., Guerberof-Arenas & Asimakoulas, 2023).

The question of *where* such competences are best acquired quickly follows. Scholars including Austermuehl (2013) and Sánchez-Castany (2023) have commented on the regrettable tendency to “silo” technology-related competences in tools-focused courses, instead of integrating them in more authentic settings and activities (e.g., Kiraly & Massey, 2019). Unfortunately, this tendency has proven difficult to reverse (cf. Section 5.3). *When* such competences should be acquired is another thorny and unresolved issue (Austermuehl, 2013; Rico & González Pastor, 2022; Sánchez-Castany, 2023; Vandaele, 2017).

Another is *to do what*, that is, what roles translator education programs should prepare graduates to play. There is increasing recognition of the diverse tasks performed by language

professionals in today's industry, including posteditor, project manager, language/cultural advisor, MT literacy consultant, and MT developer/evaluator (e.g., Angelone, 2022; Ehrensberger-Dow et al., 2023; Lehr et al., 2021). Adapting to these various roles requires graduates to acquire the wide range of competences discussed above, and to be able to adapt quickly and effectively as tasks and roles continually evolve.

2.2 The proposed MTr

After analyzing the literature and existing translation programs, and in an effort to provide respondents in our market study with a stimulus for reflection, we proposed a program outline for a 45-credit Master's program in Translation (MTr). It includes a common core of English/French transfer courses (2 general translation, 2 specialized translation, 2 in translation in a specific high-demand area of specialization) along with courses in terminology and documentation, machine translation and postediting, translation theories, bilingual revision, and professional aspects of translation (i.e., translation as a professional activity, in which themes such as continuing professional development, the role of professional associations, employment options and strategies, and ethics may be explored). Complementing this core are two additional optional, two-course modules, allowing students to focus on complementary areas such as translation in another specialized field or in another language direction or combination, computer-aided translation and terminology management, interpreting, theories, or translation as a profession. (See Appendix A for more details.) Additionally, options of a capstone project or a work placement (practicum or CO-OP) are proposed.

3 Methodology

After an initial exploration, we identified key stakeholders in the field to gather information through interviews and anonymous online questionnaires.¹ The process is described below.

3.1 Participants

We aimed to recruit representatives of various stakeholder groups: employers, professional associations, students, alumni, professors, and prospective students. Unfortunately, it was extremely difficult to recruit prospective students. However, we conducted interviews with a professional association, 4 employers, 2 students, and 3 professors, and received completed questionnaires from 84 alumni and 13 current students. (Some stakeholders belonging to more than one group are listed here in their primary affiliation.)

These participants all represent the Canadian context, in which official English-French bilingualism shapes the industry, and in which professional translation has traditionally been taught at the Bachelor's level (although several professionalizing Master's programs have recently been developed).²

¹ The research was carried out under University of Ottawa Research Ethics Board certificates S-04-23-9197 and S-06-23-9832.

² A review of translator education programs offered at Canadian universities is unfortunately beyond the scope of this paper, but some information about universities offering such programs can be found for example on the site of the Canadian Association of Schools of Translation (CAST) at <http://acet-cast.ca/>.

Although the interviews included various profiles, the student and alumni questionnaires reflect a fairly homogeneous population: most alumni originated from Ontario or Quebec, had completed professionalizing programs, had completed some postsecondary studies (CÉGEP, college or university) before entering a Translation program, and were now employed full-time in the public sector, in the National Capital Region (Ottawa/Gatineau). Most held a job closely related to their studies in translation: 31% reported that their primary job was translation, 15% were revisers, 9% project managers, 8% editors, and 7% writers. The students indicated similar intentions: over 38% wanted to be translators and 15% editors, while others expressed an interest in terminology, teaching, and other options. While only 25% of the alumni reported membership in a professional association, over half of the students intended to become members. This may be partly due to a recent change in the policy of the *Ordre des traducteurs, terminologues et interprètes agréés du Québec* (www.ottiaq.org), which now allows graduates of approved professional programs to become members on the strength of their degree.

3.2 Data collection

The entirely anonymous questionnaires were distributed via Survey Monkey (www.surveymonkey.ca). Separate, adapted questionnaires containing both closed- and open-ended questions were distributed to the stakeholder groups. The questionnaires first explored the respondents' demographic profile, and then addressed two main topics: perceptions of the language industry (e.g., employment opportunities, nature of employment, role of technologies, factors that might attract students to the industry or deter them from entering it), and opinions concerning translation programs, including the MTr proposal (e.g., important characteristics and components of translator education programs, strengths and weaknesses of the proposed program). Depending on the respondents' profiles and answers, relevant questions and options were displayed. In the alumni questionnaire, the shortest pathway consisted of a consent question, 10 demographic questions, 4 questions focusing on perceptions of the language industry, 21 questions eliciting program feedback, a question inviting them to view supplementary optional questions, and a draw entry question. Respondents who were currently employed were asked an additional 6 questions about their employment situation (e.g., role, full- or part-time status, and location), and respondents who opted in viewed an additional 17 questions about their program experience and preferences. The total number of questions thus varied from 38 to 61. Students were invited to complete an additional 5 questions about prospective employment after their programs, for a total of 43 to 66 questions. All questions, except for the consent questions, were optional. (Appendix B presents the key questions reported on in this paper.)

The semi-directed interviews began with similar questions, but evolved differently depending on the participants' observations and priorities, offering opportunities to explore important themes and priorities for each individual.

3.3 Data analysis

The questionnaires allowed for some descriptive quantitative analysis of closed-ended questions, although the relatively small sample excludes advanced statistical analysis. The main analysis was qualitative, carried out using a bottom-up thematic coding approach. The interviews were first transcribed using Microsoft Word's speech recognition function and manually edited, then coded in NVivo qualitative data analysis software (<https://lumivero.com/products/nvivo/>). The team developed a coding guide collaboratively starting from an initial analysis of important themes in the literature, and then adjusting to

better reflect the important and recurrent themes in the interviews. The codes were then also applied to the free-text answers from the questionnaires. Some examples are shown in Table 1.

Code	Example(s)
Competences	Language, Transfer, Revision, Postediting, Technology, Professional
Technologies	MT, CAT, AI
Working conditions	Employment status, Workplace, Productivity expectations
Specialization	Domain knowledge acquired/required
Remuneration	Salary, Rates
Appreciation	Appreciation for complexity and contribution of human translation

Table 18. Examples of codes used in qualitative data analysis

4 Results

In this section, we will explore participants' perceptions of the language industry, as well as the priorities identified for translator education and the strengths and weaknesses of the proposed program.

4.1 Awareness of technologies

While technologies were a focus for our study, and targeted in some questions, we did not specify particular types of technologies, allowing respondents instead to specify the technologies they found relevant for their work and for the industry. While it was not surprising that respondents commonly mentioned CAT tools and MT, it was enlightening that artificial intelligence (typically genAI) was also identified as a major influence for the industry, despite its recent advent (e.g., *I... think we are on the verge of a paradigm shift with the advent of generative AI; Generative AI is disrupting all aspects of the language industry at a rate that would outpace any university program*).³

4.2 Perceptions of the language industry

In our preliminary quantitative analysis, respondents to the student and alumni questionnaires were overall positive about the current language industry, with over two thirds (strongly) agreeing that it offered interesting employment opportunities. However, when asked if this would remain the case in the coming years, the figure dropped to just over half. Clues that help to explain this decrease may be found in the alumni answers to open-ended questions about what might deter potential students from entering translator education programs: over 40% of

³ Direct quotations taken from questionnaires and interview transcripts are shown in italics.

the 79 respondents to this question identified concerns over technologies (far beyond the 15% who mentioned employment prospects and working conditions, the two next most frequent responses). Moreover, these points may also be closely linked to technology implementation, as will be discussed below. Over 85% of the 84 alumni (strongly) agreed that technologies are currently causing significant changes in translation work but, more surprisingly to us, only two thirds (strongly) agreed that technologies would not replace translators in the near future. The fact that a third of the respondents felt doubts about the potential to replace translators certainly bears further examination in detail. On a more optimistic note, 76% of the respondents felt that the language industry offered a variety of interesting jobs, and 92% that individuals trained in translation are able to play many different roles in the language industry. This is also reflected in the diverse jobs alumni reported holding (cf. Section 3.1).

Qualitative analysis helps to reveal some reasons for respondents' concerns over technologies. These include the well-recognized fears of human translators' replacement (e.g., *[A lot of people] are asking why we have not yet been completely replaced by MT [TR]*),⁴ shifts in the nature of tasks assigned to professionals (e.g., *I... worry... that the work will shift to revising MT documents instead of actual translation*), and the adaptations required in the process (e.g., *I was passionate about "manual" translation, but... I am having trouble adapting to technologies [TR]*). Concerns were also raised about language quality in the short and long term (e.g., *I have colleagues who tell me that tools... can do miraculous things, while I find that the final product does not at all reflect the unique character of the language, and I hate to see them contribute to impoverishing the language while celebrating how much they are making per hour [TR]*).

Technology integration featured much more rarely in discussions of the attractors to the language industry for today's students and graduates. Only one of the respondents mentioned technologies in this open-ended question, noting that they can boost productivity and accelerate language processing. However, this is not to say that positive comments about technologies were lacking in the responses. In contrast to the reticence expressed by several respondents (as described above), in explaining their perceptions of the language industry, others felt that the shift in tasks would not be complete (e.g., *The future is in post-editing but also translation and adaptation of texts that don't lend themselves well to MT; Human intelligence is necessary to produce functional translations. We see this particularly in technical and creative fields [TR]; Machine translations [w]ill never be able to fully grasp the cultural context of the source text*) and that the editing task could be a stimulating and engaging one (e.g., *With technological developments, translation and editing could well become more interesting because the technology will be able to handle some of the mundane aspects of the jobs and professionals will really be able to put their talents to use*). Moreover, some noted that individuals trained in translation are well prepared to take on the task (e.g., *I think old-school trained translators would make the best post-editors*), and that their ability to achieve quality is also complemented by the skillset needed to contribute to technology improvement (*Machine translation and post-editing and computer-aided technologies need two kinds of graduates: those who use these technologies, and want to ensure a best-in-class end product, and those who will work on the ongoing development and refinement of these technologies*).

⁴ Quotations that have been translated from French are indicated by [TR]. All translations are our own.

Despite the newness of genAI tools and their influence in the field, comments addressing these tools in both questionnaires and interviews specifically evoke many of the same elements discussed above. These include recognition that some feel that the technologies can replace humans, though they believe that this is an over-simplification (e.g., *I think there's a misconception that any day now AI will take our jobs away; Jobs will be replaced by AI. However, there is still a role for trained professional translators to oversee the output and add value*). They nevertheless recognize that these tools may change language professionals' tasks (e.g., *Although not all companies will use [AI], I think it'll create a higher demand for revisers than translators*). While there was some concern over the potential challenges of maintaining a focus on quality while using such technologies (e.g., *[I]t's a reality that ... the profession is going to have to live with. ... How to adapt and postedit AI submissions, but still continue with the more traditional. ... [W]hat I'm afraid of is if we just edit AI submissions... we may lose how to translate organically*), some comments also identified potential benefits of genAI integration for language professionals specifically (e.g., *[I]f anything, language professionals will be in the best position to use [AI] tools to their advantage and optimize them; The language industry will be a source of many jobs that combine human ingenuity with artificial intelligence*).

Although technologies were rarely spontaneously identified as attractors for the language industry and translator education programs, technology-linked themes were identified. In free-text comments about the positive aspects of the field that should be highlighted for prospective students, the most commonly identified theme was versatility, generally focusing on the variety of tasks and environments in the language industry (translation, localization, revision, project management, technology management, business management), and the professional's ability to handle these tasks and find interesting opportunities as a result (e.g., *The competences and knowledge acquired in translation are transferable to many other fields [TR]*). Other responses highlighted the satisfying and valuable advisory role that language professionals may play (whether or not they work with technologies), the ability to specialize in interesting or lucrative fields or text types, and flexible and pleasant working conditions.

The diversity of roles for graduates of translation programs and the need to demonstrate adaptability over very short periods (such as that surrounding the emergence of genAI) thus stand out as important themes for translator education programs.

4.3 Program priorities and evaluation

Concerning priorities for translator education programs, including their evaluation of the importance of various proposed modules (cf. Section 2.2), alumni and students showed relatively comparable responses. Weighted averages on a scale from not at all important (1) to essential (6) showed that the alumni ranked the technology-related modules highest, with related modules including revision not far behind (Figure 1).

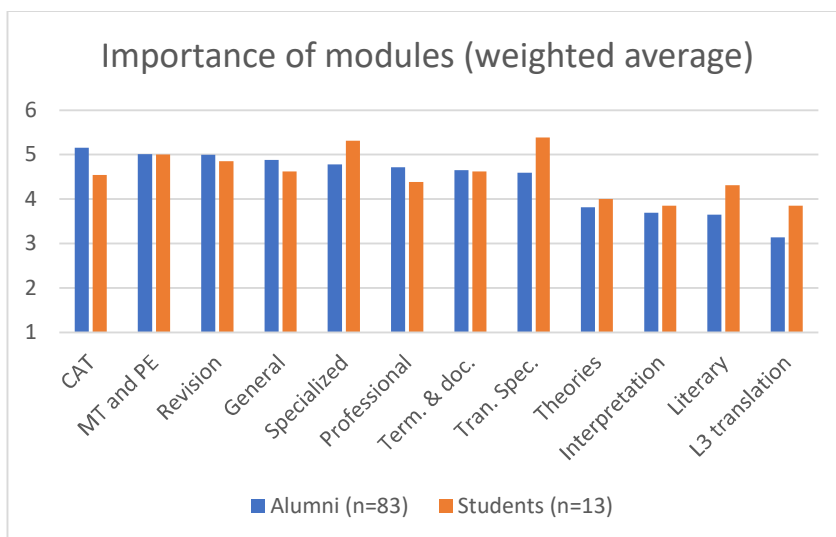


Figure 19. Weighted averages for evaluation of modules' importance

Compared to alumni, students were somewhat more focused on specialized translation courses (both the common core of specialized/technical and the specialization in high-demand fields), and on MT and postediting over CAT tools. Nevertheless, all clearly recognized the importance of technologies in training.

Complementary information in free-text comments and interviews helps to clarify priorities in technology teaching, echoing the literature (Cf. Section 2.1). Participants (primarily alumni and professors) were particularly concerned with the orientation and intended learning outcomes of technology courses, highlighting the importance of fostering critical thinking about technology (e.g., *I think it is as important to... develop students' critical thinking skills about these tools as to teach their use [TR]; [MT and postediting are] an inescapable issue that we must analyze critically. It is not enough to teach students "how to postedit." They must also discover the limits of the technologies, the debates that surround them, and the implications for [translation] [TR]; I think we have to move towards teaching them the ethics of using machine translation.... [W]e have to teach them how to understand that, well, if this is classified information, this can't go on [an online MT system].*

5 Analysis and discussion

In the data above, we can identify implications for recruitment of students into translator education programs in Canada, and derive support for several choices in curriculum design. We believe that these implications may also provide useful indications for other programs.

5.1 Implications for recruitment

As described in section 4.2, some respondents appear to have concerns about the possible replacement of human translators. (It should be noted that this may not indicate their own belief that technologies can achieve human quality, but rather perceived attitudes and priorities of decision-makers [clients, employers] who may not have the same sensitivity to the limitations of technologies.) Even if they do not feel as if replacement is a real possibility, they often recognize that perceptions of technologies may deter new students from entering the field. This

will clearly need to be addressed for reformed professionalizing translation programs to succeed. However, we believe that several other issues also require attention.

The love of languages and the written word and the satisfaction of crafting high-quality texts have long been major attractors to the language industry. Our respondents continue to assert the importance of these elements, and the resulting dedication to quality, in today's market. Nevertheless, our respondents also identified the need to highlight market requirements that would inform students' expectations of the reality they will face. Based on the feedback from our participants and our own experience, we believe that providing an accurate picture of what graduates can expect in the language industry will entail a considerable adjustment of the messages that have long been central to recruitment of new language professionals, just as the programs will need to evolve.

As our respondents indicated, not only are there opportunities in (slightly different but nevertheless language-focused) occupations, but also in new and emerging roles being created by technology implementation. Respondents highlighted diverse tasks and roles in today's language industry and the fact that translator education programs are already preparing students to adapt. The range of occupations held by alumni supports this assertion. We agree that such versatility and agility constitute key qualities of future language professionals that we must seek—and then cultivate—in students admitted to translator education programs. As one of our respondents put it, *[J]udgement and versatility are... important... [because] it's what sets us apart from the machine.* The element of judgment, also evoked in many of our other respondents' comments, is identified as one of the guarantors of continued employment opportunities, handling tasks for which technologies alone are not sufficient, as well as helping to develop, evaluate, choose, implement, and monitor these technologies. These existing and emerging occupations share a common element of agency, a concept that has been a fruitful one for discussions of the influence of technologies on the industry from a sociological perspective (e.g., Olohan, 2011; Ruokonen & Koskinen, 2017), and one that was evoked by many of our respondents. As agency can be seen as a means of empowerment, we feel that highlighting it when describing the language industry can help to reassure aspiring language professionals that their work will be stimulating and rewarding. (As one of our participants put it, *The translator is at the centre of the translation process and uses the tools they master like the conductor of an orchestra [TR].*)

While the translator's role has often been seen as solitary, our respondents helped to highlight the increasing opportunities for collaboration and teamwork in more technologized workflows. Moreover, many highlighted the advisory role that language professionals may play in ensuring that translation quality is maintained, technologies are used effectively and responsibly, the needs of society and of cultural and language communities are respected, and the true worth of language professions is recognized. This may involve an educational role in interactions with colleagues, employers and clients, as well as the general public (Ehrensberger-Dow et al., 2023; Lehr et al., 2021).

5.2 Implications for curriculum design

Of course, promising preparation for the new reality of the language industry and its workplaces is not enough; programs must then deliver on that promise. This has various implications for curriculum design.

One key tool in preparing students for the realities of the working world is authentic workplace learning opportunities (e.g., CO-OP or practicums). These are not only highly

regarded by students, alumni and employers in the study; some element of authentic, supervised work is required for the recognition of professionalizing programs for the professional association OTTIAQ and thus for membership in the association on the strength of the degree. Certification is also of considerable interest to the students who completed our questionnaire. Thus, the inclusion of work experience in the program becomes unquestionable.

The importance of both technologies and critical thinking about them also has important implications for curriculum design. One essential consideration is that, while of course technology-focused courses are important in this type of program, it is neither advisable nor even possible—while still reflecting the authentic experience prized by employers, alumni and students—to restrict technologies to dedicated courses (cf. Section 2.1). Rather, we assert that technology-related skills, and thus critical thinking about technologies, should be integrated across a range of courses in ways that complement and enrich other skills being acquired. Figure 2 illustrates some of our ideas for integrating various elements of critical thinking about technologies (CAT, MT and genAI) into the various modules proposed in the MTr.

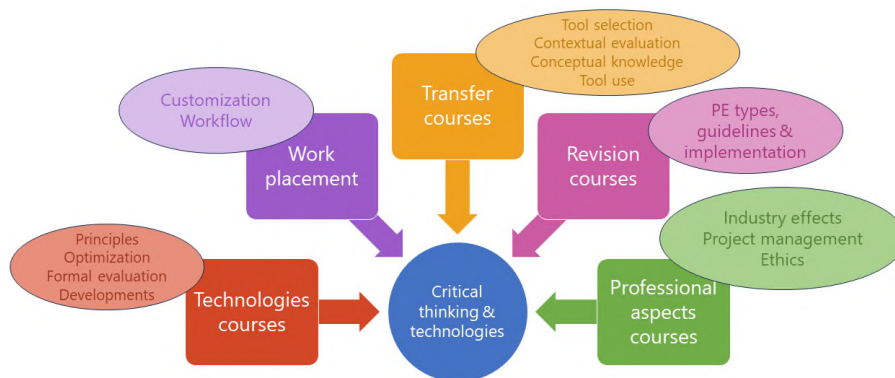


Figure 2. Integration of critical thinking about technologies in various program elements

The data support our belief that it is necessary to ensure that students are aware of the basic principles behind various technologies and how these principles affect tools’ functioning, potential and limitations, as well as of new directions being explored for future development. We would add that they should also be introduced to various strategies for optimizing and evaluating technologies’ performance within those limitations. All of these slot nicely into courses focussing specifically on technologies, along with a basic introduction to tool use. However, we must stress that implementation of the tools in transfer (translation) courses is an essential complement to the technology courses, to ensure that students practice using technologies for genuine translation tasks, as well as choosing the right tool for the right job and evaluating its performance in context appropriately. Moreover, the research skills and linguistic and conceptual knowledge acquired in courses in specialized subject fields will be valuable in assisting students in evaluating performance, identifying problems (e.g., inappropriate or inconsistent terminological choices from MT, hallucinations from genAI), and improving on suggestions from tools to achieve the required quality for a given translation brief. Revision courses offer precious opportunities to explore distinctions between revising human translation and postediting (cf. do Carmo & Moorkens, 2020) and various types of postediting and the associated guidelines, as well as to develop and practice postediting skills. The negotiations and interactions between stakeholders and technologies in complex workflows, as well as the experience of working with customized technologies (e.g.,

customized MT systems, large termbases and translation memories), may be challenging to recreate in an academic setting (Koponen et al., 2023). However, work placements (CO-OP, practicums) can allow students to experience authentic workflows on a realistic scale. Finally, and just as importantly for encouraging critical thinking, professional aspects courses provide opportunities for students to reflect on the effects of technologies on the industry and to learn some of the skills that are required to integrate technologies appropriately and responsibly, as well as to examine ethical implications of technology use in various contexts. Like many specialists in translation technology pedagogy, we firmly believe that this kind of integration will be essential to meet the expectations and demands of today's industry. Nevertheless, we recognize the challenges of ensuring sufficient preparation for both "purely" human and technology-augmented translation, while remaining within the desirable length for a program of this type (e.g., Austermuehl, 2013; Koponen et al., 2023).

5.3 Challenges of technology integration

As observed in previous work at the University of Ottawa (e.g., Bowker & Marshman, 2010; Marshman & Bowker, 2012) and in many other programs (e.g., Austermuehl, 2013; Rico & González Pastor, 2022; Sánchez-Castany, 2023), while certainly worthwhile, integrating technologies across the program is not straightforward. First and foremost, it requires effective collaboration between faculty members to ensure that tool choices are as coordinated as possible, providing a good balance of consistency and variety, as well as completeness without excessive repetition. Even once balance is achieved, quickly evolving technologies require that program content (and thus, teaching and learning resources such as tutorials)⁵ be constantly updated to adjust to new realities. As noted in the introduction, preparing for constant adaptation at a pace that far exceeds that of the typical university program revision also requires careful planning of courses and descriptions to provide as much flexibility as possible while maintaining an appropriate standard. Particularly in the case of a Master's degree, there is an added challenge of adapting to students' previous experience and training. Moreover, the timing with which various elements are introduced has been a matter of some debate in the field of translation pedagogy (cf. Section 2.1), and no clear answer has emerged from our discussions. While several participants expressed concerns about the balance between "purely" human and technology-augmented translation skills in translator education, with some encouraging the acquisition of skills in translation "from scratch" before integrating technologies, no unequivocal answer emerged as to the right skills to introduce at a given time in the program. More reflection—and no doubt experience—will be required to clarify this question.

6 Concluding remarks

The strong focus on technologies in the participants' contributions, even in more generally oriented questions, identifies technology training as a central theme for future program development. The speed with which newer technologies such as genAI have been recognized as essential in translator education demonstrates that programs must strive not only to keep up with technology development, but also to prepare graduates to adapt to—and make the best

⁵ The approach described by Vieira et al. (2021) offers advantages from this perspective, as in addition to fostering student autonomy it also reduces the need to provide detailed resources for learners. Of course, it also entails its own challenges, as the authors describe.

of—changes yet to come. In line with the literature, employers, alumni and students have recognized the importance of coming to grips with technologies not only at a practical level, but also at the level of reflection and critical evaluation. Future programs must address how technologies continue to transform employment opportunities, tasks and workflows, and their potential effects on the future of the profession and the well-being and satisfaction of professionals. Translator education must prepare graduates to be agile and adaptable and to embrace the evolution of technologies throughout their careers, while maintaining a high level of human skill, including judgment, critical thinking, and creativity. Integrating technologies throughout translator education programs, in authentic contexts and workflows, will provide diverse opportunities to develop these capacities. We hope that this will continue to prepare graduates to play a wide range of roles in the language industry as employment prospects evolve and diversify.

Our data reveal a sense of uncertainty and instability—but also of promise—resulting from technology evolution. While there are certainly concerns, and some respondents even feel as if they may lose their jobs to technologies, there is also a sense of positivity among many current professionals. Moreover, employers have very clearly highlighted a need for promising new recruits to join the ranks of language professionals and meet the changing needs. By better preparing future students for the realities they will face, we hope to ensure that graduates are ready not only to succeed in the new reality of the language industry, but to flourish.

Acknowledgements

Thanks are extended to all participants in the market study for their valuable contributions, as well as to the colleagues at the School of Translation and Interpretation for their contributions to the project and the Faculty of Arts of the University of Ottawa for their support.

References

- Angelone, Erik. 2022. Weaving adaptive expertise into translator training. In Gary Massey, Elsa Huertas-Barros, & David Katan, editors, *The Human Translator in the 2020s*. Routledge, London, pages 60–73.
- Austermuehl, Frank. 2013. Future (and not-so-future) trends in the teaching of translation technology. *Tradumàtica: Tecnologies de la Traducció*, 11. <https://doi.org/10.5565/rev/tradumatica.46>
- Bowker, Lynne 2020. Translation technology and ethics. In Kaisa Koskinen and Nike K. Pokorn, editors, *The Routledge Handbook of Translation and Ethics*. Routledge, London, pages 262–278.
- Bowker, Lynne, and Elizabeth Marshman. 2010. Towards a model of active and situated learning in the teaching of computer-aided translation: Introducing the CERTT project. *The Journal of Translation*, 13(1/2): 199–226.
- do Carmo, Félix, and Joss Moorkens. 2020. Differentiating Editing, Post-Editing and Revision. In Maarit Koponen, Brian Mossop, Isabelle Robert and Giovanna Scocchera, editors, *Translation Revision and Post-editing*. Routledge, London, pages 35–49.
- Ehrensberger-Dow, Maureen, Alice Delorme Benites, and Caroline Lehr. 2023. A new role for translators and trainers: MT literacy consultants. *The Interpreter and Translator Trainer*, 17(3): 393–411. <https://doi.org/10.1080/1750399X.2023.2237328>

- European Master's in Translation Group. 2022. *European Master's in Translation Competence Framework 2022*. European Master's in Translation Group. https://commission.europa.eu/system/files/2022-11/emt_competence_fwk_2022_en.pdf [last accessed 31 January 2024]
- Guerberof-Arenas, Ana, and Dimitris Asimakoulas. 2023. Creative skills development: Training translators to write in the era of AI. *HERMES - Journal of Language and Communication in Business*, 63. <https://doi.org/10.7146/hjlc.vi63.143078>
- Kiraly, Donald C., and Gary Massey, editors. 2019. *Towards Authentic Experiential Learning in Translator Education* (2nd edition). Cambridge Scholars Publishing, Newcastle upon Tyne.
- Koponen, Maarit, Alan Melby, and Amina Tahraoui. 2023. Educating the next generation of translators in the age of AI. Paper presented at Translating and the Computer - TC45, Luxembourg.
- Lehr, Caroline, Maureen Ehrensberger-Dow and Alice Delorme Benites. 2021. MT literacy as a means of agency and empowerment for translators. Paper presented at the Łódź-ZHAW Duo Colloquium on Translation and Meaning, online, Winterthur.
- Marshman, Elizabeth, and Lynne Bowker. 2012. Translation technologies as seen through the eyes of educators and students: Harmonizing views with the help of a centralized teaching and learning resource. In Séverine Hubscher-Davidson and Michal Borodo, editors, *Global Trends in Translator and Interpreter Training: Mediation and Culture*, Bloomsbury Academic, New York, pages 69–95.
- Massey, Gary, and Maureen Ehrensberger-Dow. 2017. Machine learning: Implications for translator education. *Lebende Sprachen*, 62(2): 300–312. <https://doi.org/10.1515/les-2017-0021>
- Mellinger, Christopher. D. 2017. Translators and machine translation: Knowledge and skills gaps in translator pedagogy. *The Interpreter and Translator Trainer*, 11(4): 280–293. <https://doi.org/10.1080/1750399X.2017.1359760>
- Moniz, Helena, and Carla Parra Escartín, editors. 2023. *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*. Springer International Publishing, Cham. <https://doi.org/10.1007/978-3-031-14689-3>
- Moorkens, Joss. 2022. Incorporating ethics in translation programmes. *European Master's in Translation Network Blog*. <https://blogs.ec.europa.eu/emt/incorporating-ethics-in-translation-programmes/> [last accessed 31 January 2024]
- Olohan, Maeve. 2011. Translators and translation technology: The dance of agency. *Translation Studies*, 4(3): 342–357. <https://doi.org/10.1080/14781700.2011.589656>
- Pierce, Rachel. 2023. 5 Tedious Non-Translation Tasks ChatGPT Can Do Amazingly Well. *American Translators Association (ATA)*. <https://www.atanet.org/business-strategies/5-tedious-non-translation-tasks-chatgpt-can-do-amazingly-well/> [last accessed 6 July 2023]
- Pym, Anthony. 2023, April 27. *GPT in the training of translators*. https://www.youtube.com/watch?v=b9U_FUaneso [last accessed 31 January 2024]

- Rico, Celia, and Diana González Pastor. 2022. The role of machine translation in translation education: A thematic analysis of translator educators' beliefs. *Translation & Interpreting*, 14(1): 177–197. <https://doi.org/10.12807/ti.114201.2022.a010>
- Rodríguez-Castro, Monica. 2018. An integrated curricular design for computer-assisted translation tools: Developing technical expertise. *The Interpreter and Translator Trainer*, 12(4): 355–374. <https://doi.org/10.1080/1750399X.2018.1502007>
- Roy, Jean-Hugues, and Éric Poirier. 2023. *La traduction a survécu à l'IA. D'autres métiers qui semblent menacés par ChatGPT survivront aussi*. La Conversation. <http://theconversation.com/la-traduction-a-survecu-a-lia-dautres-metiers-qui-semblent-menaces-par-chatgpt-survivront-aussi-200413> [last accessed 31 January 2024]
- Ruokonen, Minna, and Kaisa Koskinen. 2017. Dancing with technology: Translators' narratives on the dance of human and machinic agency in translation work. *The Translator*, 23(3): 310–323. <https://doi.org/10.1080/13556509.2017.1301846>
- Sánchez-Castany, Roser. 2023. Teaching translation technologies: An analysis of a corpus of syllabi. In Gary Massey, Elsa Huertas-Barros, and David Katan, editors, *The Human Translator in the 2020s*, Routledge, London, pages 27–43.
- van der Meer, Jaap. 2023. ChatGPT (Un-)Fit for Translation? *TAUS*. <https://www.taus.net/resources/blog/chatgpt-un-fit-for-translation> [last accessed 31 January 2024]
- Vandaele, Sylvie. 2017. L'étudiant, l'enseignant et la technologie. *Hermēneus. Revista de traducción e interpretación*, 19. <https://doi.org/10.24197/her.19.2017.1-17>
- Vieira, Lucas Nunes, Xiaochun Zhang and Guoxing Yu. 2021. 'Click next': On the merits of more student autonomy and less direct instruction in CAT teaching. *The Interpreter and Translator Trainer*, 15(4): 411–429. <https://doi.org/10.1080/1750399X.2021.1891515>
- Yamada, Masaru. 2023a. Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability (arXiv:2308.01391). arXiv. <https://doi.org/10.48550/arXiv.2308.01391>
- Yamada, Masaru. 2023b. Enhancing Translation Accuracy and Efficiency through ChatGPT. CTS Convergence Lecture, Virtual.

Appendix A: Optional course modules

In addition to core courses, the optional courses below are proposed.

- **L2 translation**
 - Two translation courses in the non-dominant language direction
- **Specialization in Translation**
 - Two translation courses in another specialized field
- **Terminology and Computer-aided translation**
 - Terminology management in professional translation
 - Computer-aided translation and localization
- **Translation theories**

- Comparative stylistics
- Translation and interculturality
- **Translation as a profession: 2 of**
 - Marketing in the language industry
 - Small business management in the language industry
 - Project management in the language industry)
- **Translation in/to a third language (L3)**
 - General translation to and from a third language
 - Specialized translation to and from a third language
- **Interpreting**
 - Sight translation
 - Introduction to interpreting
- **Literary translation**
 - Translation and literature
 - Literary translation workshop

Appendix B: Key questions from the questionnaires

Perceptions of the language industry (Alumni/Students)⁶

Matrix question: Please share your perceptions of the language industry and the employment opportunities it offers.

Answer options: Strongly disagree, Disagree, Neutral, Agree, Strongly agree, I don't know/not applicable

Statements:

I like my current job in the language industry.

My work situation is stable and secure.

The language industry offers interesting employment opportunities.

In the coming years, the language industry will offer interesting employment opportunities.

Translation is still an interesting career.

Technologies will not in the near future replace human translators.

Technologies are currently causing significant changes in translation work.

Individuals trained in translation are able to play many roles in the language industry.

The language industry offers many different types of interesting jobs.

⁶ Statements included only in the alumni questionnaire are shown in italics.

The language industry currently offers good salaries.

If I were starting my degree today, I would study Translation.

When I am asked for career advice, I recommend translation.

Free-text questions

If you would like to explain or comment on your answer above, please do so here.

What would you identify as the main advantages of a career in the language industry that we should promote to potential translation students?

What are the main negative perceptions of a career in the language industry that need to be addressed to encourage potential students to study translation?

Program evaluation (Alumni/Students)

Matrix question: Below, you will see a list of (compulsory and optional) modules proposed as part of our professional Master's in Translation. Please indicate what you think of each of these.

Answer options: Not at all important, Not very important, Neutral, Important, Very important, Essential, I don't know

Modules:⁷

General translation

Specialized and/or technical translation

Specialization in translation (translation in a specialized field)

Terminology and documentary research

Machine translation and postediting

Professional aspects of translation

Bilingual revision

Translation theories

Translation in/to a third language (e.g., Spanish)

Computer-aided translation technologies

Literary translation

⁷ Modules in which at least one course is compulsory are underlined. In some areas (e.g., professional aspects, translation specialization, translation theories), additional optional modules can be added to the compulsory course(s).

Interpreting

Free-text questions

If you would like to comment on your responses above, please do so here.

Do you see any particular strengths of the proposed program? Please share them here.

Do you see any particular weaknesses of the proposed program? Please share them here.

Studying the Need to Optimize Search for Amendments and Corrigenda in EU Institutional Translation

Timea Palotai-Torzsas

Juremy Ltd.

timea@juremy.com

Robin Palotai

Juremy Ltd.

robinp@juremy.com

Abstract

To comply with EU institutional translation standards, linguists must carefully search for amendments and corrigenda to ensure accuracy and consistency. Our study explores the importance of consolidation: the action of combining an initial act with its subsequent amendments and corrections in a single consolidated document. Firstly, we discuss specific translation scenarios where it is critical to consult consolidated documents, as well as corrigenda and amendments not covered by consolidation, and highlight the challenges they present. Secondly, we provide statistics on the proportion of documents affected by modifications and/or consolidation in a fundamental segment of the EU legislative corpus. We examine the set of regulations, directives, and decisions adopted as basic acts by the ordinary legislative procedure, drawing on statistics on the extent of consolidation, as well as on unincorporated amendments and corrigenda. We found that the majority of the regulations and directives examined have a consolidated version, and that non-consolidated modifications in this segment are rare. Our results underline the need for careful and laborious research on the history of reference documents and their metadata. We aim to improve this process through our online concordance tool Juremy.com, by displaying metadata on consolidation and corrigenda, and thereby further support linguists in achieving high translation quality.

1 Introduction

To comply with EU translation standards on consistency and accuracy (Stefaniak, 2017), linguists search the EU corpus thoroughly to find the equivalent of a given term or phrase that has already been translated into their target language. This task might seem straightforward when the source phrase is included in a document already published in the Official Journal of the EU, translated into the target language as one of the 24 official languages.

However, finding the source language expression in an initial legal act or case-law, and matching it with its corresponding target language variant is not always sufficient to provide a correct translation. Linguists need to take an extra step to ensure accuracy: they also need to check whether the reference document has been subject to consolidation and/or modifications (corrigenda or amendments).

In our study, we examine why consolidated documents are a significant resource from the translator's perspective. We also demonstrate the importance of taking into account corrigenda and amendments in the EU translation process by providing statistics on the proportion of consolidated and corrected documents in a fundamental segment of the EU legislative corpus.

2 Context

Consolidation involves “the action of combining an initial act and all its subsequent amendments and corrections in a single document.” Although they have no legal effect, consolidated texts, are important technical documents as “they show the legal rules that are applicable at a certain point in time” (EUR-Lex, 2023a).

In several translation scenarios, it is crucial to research the amendment history and existence of corrigenda of the reference document:

One example is when the source text – mostly a judicial text – refers to regulations which were applicable at an earlier period of time but are no longer in force due to amendments or the repeal of the legal act in question. In such cases, the translator must search for the version of the reference text applicable at that specific time in the past (Kokkinidou and Giovani et al., 2023). For this purpose, consolidated versions are a particularly useful resource.

Another reason to consult consolidated versions is that they cover (most of the) corrigenda related to the basic act. A corrigendum is “an instrument published in the Official Journal which formally rectifies an error in one or more language variants of an EU legal document” (Biel and Pytel, 2020). Corrigenda published in relation to the legal act must be taken into account when choosing the right terminology. Careful research is required as in some cases, corrigenda are published years after the publication of the initial version of the act (Bobek, 2009; Biel and Pytel, 2020).

In contrast to amending acts, which are available in all official EU languages, corrigenda often affect only one or a few language variants of a document. This is the case when a corrigendum is not a “source-text corrigendum” (Biel and Pytel, 2020), rectifying errors in the English version, but a correction of errors in a translation in a given target language. In this case, the bilingual view of the consolidated document might not be available in the desired language pair.

3 The problem

Imagine that you are a translator and must translate a text containing a reference to a legislative document in the EU corpus. Moreover, you are mandated to strictly adhere to the terminology previously used. Assuming that while translating a given source segment, you find a perfect or fuzzy match hit from a basic legislative act – for example using your CAT tool’s local translation memory, or using Juremy’s fast full-corpus search. You also determine that the domain and context of this basic act fits your topic quite well.

At this point, should you trust the target segment found as-is, or should you continue your search? This is the situation motivating our analysis.

If we omit further searching, we risk missing potential corrigenda that retroactively correct either the source and/or the target segment. In the case of a modified source segment, it is situation-dependent whether the non-modified original can be considered as valid to use as a reference for translation. We are more concerned about the case of a corrected target segment, where using the original target as-is would very likely lead to incorrect terminology.

3.1 A quick look on the corrigendum research procedure

To better understand the corrigendum search procedure, we interviewed Juremy users and also examined the search process ourselves in greater detail. Based on this, the following steps illustrate the post-hit searching which can be performed using EUR-Lex:

Open the reference document containing the source segment match on EUR-Lex. Make sure that the display language is set to your target language.

Then, open the latest consolidated version affecting your target language (if listed), and also the Document Information page of the base document, in new browser tabs.

Examine the preamble of the consolidated version for the C1, C2, ... consolidated corrigendum markings, and jump to them, or search for them in the consolidated document, to see if they affect the part of the document you are currently translating. This is effective if there are few and compact corrigenda, but not otherwise.

Alternatively, you can open the consolidated document in the bilingual view and search for your source segment. But this can be problematic if the only consolidated modifications are corrigenda which are available only in a few languages (and not your source language)", because then the consolidated document will not be available for the bilingual view in your language pair.

If there is no consolidated version, or the search based on this is not conclusive, switch to the tab containing the document information, scroll to the modifying documents section, and look for listed corrigenda. You will probably need to open the ones indicated to affect your source and/or target language and inspect the corrigendum text itself.

3.2 Our questions

From the above description we can see that corrigendum research is quite laborious. But failing to perform it or missing out some steps can jeopardize the correct use of terminology, and even undermine consistency with other references within the same text (where the corrected version of the term is used).

Therefore, on the one hand, we seek technological measures which EUR-Lex, or EU-terminology specific translation-assisting software like Juremy, can employ to speed up corrigendum research. While, on the other hand, we attempt to determine how impactful these measures would be by quantifying:

- How many legislative acts are corrected?
- How many corrected documents are consolidated?
- How many corrigenda are not covered by consolidation?
- The prevalence of all-language corrigenda (that is presumably source-text)?
- How many language variants are usually affected by corrigenda?

4 Methodology

We examine the three types of basic acts in Celex sector 3 (legal acts) adopted by the ordinary legislative procedure (OLP): regulations, directives and decisions. By basic act we mean legal acts which can serve as a basis for consolidation according to the EU's consolidation methodology (EUR-Lex, 2022), but which do not necessarily have a consolidated version.

We focus on the OLP documents in our initial research because of their core importance. We expect OLP documents to be consolidated with priority, therefore the modifications not yet covered by consolidation should approximate a lower bound – or optimistic scenario – over the full set of documents.

We draw statistics on the extent of (1) consolidation, (2) amendments and corrigenda not yet incorporated into the latest consolidated version, and (3) amendments and corrigenda in non-consolidated acts.

4.1 Document metadata

Comprehensive metadata are collected for EUR-Lex documents in sector 0 (consolidations), 1 (treaties) and 3 (legislation) by querying the EUR-Lex webservice. Then this metadata is postprocessed to establish document relations and features on which we base our statistics.

4.2 Legal acts adopted by the ordinary legal procedure

We treat a sector-3 document as OLP-adopted if metadata indicates creating agents of EP and CONSIL. We note that this does not fully coincide with the official EUR-Lex website's statistics filter for OLP (EUR-Lex, 2023b), most notably we do not exclude some 200 budgetary procedure documents. Performing this is subject to future research.

4.3 Modifications

We distinguish two kinds of modifications, corrigenda and amendments. We recognize corrigenda based on their Celex identifier structure. We treat a document as amending if its metadata includes either the 'AMENDS' and 'ADDS TO' document links. We do not treat a document as amending if it just contains 'REPEALS' document linkage.

4.4 Basic acts

We treat an act as basic if it is not a corrigendum, and also if it does not amend any other document, or if it has a consolidated version. We assume that having a consolidated version indicates that the document is of a certain importance in itself, making it more likely that it has its own standalone provisions. While this appears to be mostly true based on manual sampling, it is not universally true.

Another option would be to rely on the 'REP' directory code, on which the above mentioned EUR-Lex statistics filter bases its basic or amending categorization, and which is a manually applied label (not without false positives either).

As a result of our divergences from the EUR-Lex statistics query, we treat more OLP acts as basic (1446 vs 1777, respectively). We do not believe this changes the fundamental shape of the resulting statistics.

4.5 Date range of included documents

We considered limiting the scope of examined documents to those published after the date when the current 24 languages became official, to reflect current trends better. But after running these statistics, there were no major characteristic differences at least from our perspective, so we continue to perform the analysis on the full document set regardless of publication date.

5 Statistics

5.1 Existence of corrigenda

Table 20 shows the OLP basic acts with modifications analysed by us, and breaks them down into whether they are consolidated or not. We can see that the majority (80 to 90%) of directives and regulations are consolidated, while the minority (18%) of decisions are consolidated.

	Consolidated	Non-consolidated	<i>Total</i>
Directive	475 (88%)	61 (11%)	536
Regulation	596 (80%)	143 (19%)	739
Decision	87 (18%)	389 (81%)	476
Other	-	26 (100%)	26

Table 20. Number of OLP basic acts by consolidation state.

We were interested to see whether the non-consolidated acts had modifications, or had been modified but not yet consolidated. Table 21 indicates that those acts which are not consolidated are in most of the cases also unmodified, so consolidation would not be necessary. We can also see that 8 to 10% of non-consolidated directives and regulations have corrigenda, not a negligible ratio.

	Unmodified only	Amended only	Corrected only	Amended & corrected	<i>Total</i>
Directive	54 (88%)	1 (1%)	5 (8%)	1 (1%)	61
Regulation	123 (86%)	5 (3%)	15 (10%)	-	143
Decision	372 (95%)	8 (2%)	7 (1%)	2 (1%)	389
Other	19 (73%)	-	7 (26%)	-	26

Table 21. Number of modification types affecting non-consolidated OLP basic acts.¹

Turning to the consolidated acts, Table 22 shows that about 85% of consolidated directives and regulations are affected by corrigenda, and 36-40% are affected by corrigenda alone (no amendments). In the case of decisions, the proportion of corrigenda and amendments are more balanced: 49% of this type of legal act are affected by corrigenda, and 24% are only corrected. This last category of corrigenda-only consolidations are more prone to missing certain languages in the consolidated version.

¹Most of the non-consolidated acts were not modified. Some examples of non-consolidated acts which have modifications, by their Celex identifiers: 32019R0501 has an amendment, and 32023R1231 has a corrigendum.

	Unmodified	Amended only	Corrected only	Amended & corrected	<i>Total</i>
Directive	1 (1%)	71 (14%)	192 (40%)	211 (44%)	475
Regulation	-	86 (14%)	217 (36%)	293 (49%)	596
Decision	1 (1%)	45 (51%)	21 (24%)	20 (22%)	87

Table 22. Number of modification types affecting consolidated OLP basic acts.²

To what extent do the consolidated versions cover all the existing modifications? Table 23 shows that while about three quarters of consolidated directives and regulations are fully consolidated, there are about 15% which still have unconsolidated corrigenda.

	Fully consolidated	Has uncovered amendment	Has uncovered corrigendum	Has uncovered am. & corr.	<i>Total</i>
Directive	353 (74%)	47 (9%)	63 (13%)	12 (2%)	475
Regulation	459 (77%)	47 (7%)	71 (11%)	19 (3%)	596
Decision	78 (89%)	7 (8%)	2 (2%)	-	87

Table 23. Coverage of modifications in consolidated OLP basic acts.³

²As expected, consolidated acts are heavily affected by modifications. Also, in a significant fraction of cases (“Corrected only”), corrigenda are the sole reason for consolidation. Note: the appearance of seemingly unmodified acts is due to delayed repeals that also contain temporal modifications, which we do not treat as modifying [Comment: we deleted the suggestion “modification” because the thing we wanted to point out here is the fact that the act in question is modifying another act, so it is a “modifying act”.] during our processing, as the modification is not apparent from metadata. For example, 32002L0003 was repealed by 32008L0050, but also transitionally modified by it.

³Manually inspecting a sample of the non-consolidated modifications, we observed that most often the lack of consolidation is due to a very recent modification – for example 32016L2341 modified by the recent 32022L2556 –, or an abandoned consolidation process of a repealed act – for example 32004R0808 whose amendment 32019R1700 was never consolidated –, or amendments whose effects will only start to apply in the future – for example 32012R0978’s amendment 32021R2127 will only apply in 2025. In rare cases, the metadata also fails to indicate that a given corrigendum was consolidated, while it is apparent from the consolidated text itself that it was – for example 32009R1073R(04). An interesting example that lacks consolidation of a corrigendum, regulation 32020R0852’s 32020R0852R(05) affects the preamble, but the preamble itself is not consolidated, so corrigendum is not consolidated either. Another one is directive 32011L0065, which has a recent Danish corrigendum 32011L0065R(07) not yet incorporated.

We conclude that 10 to 15% of directives and regulations have non-consolidated corrigenda. Put another way, we can expect to find one in every eight reference documents to have non-consolidated corrigenda. Also, about 60% of directives and regulations⁴ have corrigenda that are already consolidated.

5.2 Language variants affected by corrigenda

Now that we have some insight into the extent of corrections, we are interested to know how many language variants are typically affected by corrigenda. Table 24 shows that only 3 to 5% of corrigenda affect all languages of the base document. This is an upper bound to the extent of source-language corrigenda, because not every corrigendum affecting all languages is a source language corrigendum.

	All languages	Not all languages	<i>Total</i>
Directive	48 (3%)	1273 (96%)	1321
Regulation	112 (5%)	1771 (94%)	1883
Decision	2 (2%)	76 (97%)	78
Other	3 (42%)	4 (57%)	7

*Table 24. Number of OLP corrigenda affecting all language variants of documents in the various legislative categories.*⁵

So how many language variants do target-language corrigenda affect? Table 25 shows that about 70% of corrigenda affect only a single language, and another 10% affects two to three languages.⁶

⁴The 70 to 75% results from multiplying the ratio of consolidated directives and regulations (80 to 90%) by the ratio of existing corrigenda in these consolidated acts minus the ratio of uncovered corrigenda in them (85% - 15%).

⁵The corrigendum 32020R0852R(05) is a good example of a corrigendum affecting all-languages, which is also a source-language corrigendum. The corrigendum 32009R1073R(05) is also all-language, but for example its German variant contains additional corrections. 32009R0810R(01) affects many, but not all languages, and is an example of multiple, mostly independent language-variant corrections batched together into single corrigendum.

⁶We expect these latter corrigenda affecting only a few languages to be independent corrections as well, similar in spirit to single-language corrigenda, but technically issued together.

	Single language	2 to 3	4 to 10	11 to 20	21 to 24	Total
Directive	961 (72%)	150 (11%)	72 (5%)	74 (5%)	64 (4%)	1321
Regulation	1402 (74%)	184 (9%)	87 (4%)	67 (3%)	143 (7%)	1883
Decision	53 (67%)	3 (3%)	3 (3%)	16 (20%)	3 (3%)	78
Other	3 (42%)	-	1 (14%)	1 (14%)	2 (28%)	7

Table 25. Distribution of the number of published language variants of OLP corrigenda in the various legislative categories.

Therefore, we can conclude that the majority of corrigenda fall into the category of harder-to-search target-language corrigenda, instead of the source-language corrigenda.

6 Results and implications

Our results show that the majority of the examined regulations and directives, and a significant minority of the decisions examined are consolidated. On the other hand, we also find that in the case of both consolidated and non-consolidated OLP basic acts, a non-negligible portion (about 10% to 15%) of these documents are affected by non-consolidated corrigenda. Another interesting result of this study is that only a significantly small percentage of corrigenda affect all official EU languages, which implies that in the vast majority of cases in the corpus examined, elaborate research is needed to find the correct terminology based on the target-language reference text, as the bilingual view of the source and target text versions of the document will not necessarily indicate the most recent modifications of the referred document.

Our study demonstrates the importance of searching for amendments and corrigenda to comply with the quality standards of EU institutional translation. However, this is laborious and time-consuming, as linguists need to track down various metadata and document versions to find the correct applicable version of a target language text.

On the EUR-Lex website's user interface, we found that adding an indication of the languages that are affected by corrigenda to the document content view might help to shorten the search in the case where no corrigendum exists.

As developers of Juremy.com, an online concordance tool providing fast phrase-based bilingual search on the EU corpus and IATE in all 24 EU languages, we aim to enhance our service based on these findings. Indicating whether a target hit is affected by consolidation or corrigenda would further support the EU translation workflow of our users by reducing the time spent on corpus searching, and thus improving translation quality.

7 Open questions and future research

It should be taken into consideration that the scope of documents examined are the most commonly referenced and frequently consulted acts in the EU corpus, and results for the remainder of the corpus might be different for the following reasons: on the one hand, intense review during the drafting process might lead to fewer errors, but on the other hand, these frequently cited legal acts are more likely to be affected by corrigenda due to the number of

eyes looking at them post-publication. Therefore, analysing the extent of corrigenda on a broader set of documents could be revealing.

In our analysis we relied only on document metadata, and did not take document content into account. The latter might facilitate more accurate categorization of corrigenda, for example to differentiate accidental all-language corrigenda from truly source-language corrigenda.

References

- Biel, Łucja and Pytel, Izabela. 2020. Corrigenda of EU Legislative Acts as an Indicator of Quality Assurance Failures: A Micro-diachronic Analysis of Errors Rectified in the Polish Corrigenda. In *Institutional Translation and Interpreting*, pages 150-173. Routledge.
- Bobek, Michal. 2009. Corrigenda in the Official Journal of the European Union: Community Law as Quicksand. *European Law Review*, vol. 34, pages 950–962.
- EUR-Lex. 2022. Consolidation methodology. <https://eur-lex.europa.eu/content/intro/collection/Methodology-on-Consolidation.pdf> [last accessed August 2023].
- EUR-Lex. 2023a. Consolidated texts. <https://eur-lex.europa.eu/collection/eu-law/consleg.html> [last accessed August 2023].
- EUR-Lex. 2023b. Legal acts statistics. <https://eur-lex.europa.eu/statistics/legislative-acts-statistics.html> [last accessed August 2023].
- Kokkinidou, Anna, Giovani, Afroditi, and Krimpas, Panagiotis G. 2023. From lawyer-linguist to lawyer-post-editor: A changing landscape in legal translation. Conference presentation, EULITA 2023.
- Stefaniak, Karolina. 2017. Terminology work in the European Commission: Ensuring high-quality translation in a multilingual environment. Book chapter in *Quality aspects in institutional translation*, vol. 8, page 109.

Term Translation: Convert or Converse?

Aida Kostikova

Bielefeld University

aida.kostikova@uni-bielefeld.de

Kristin Migdisi

CrossLang

Sara Szoc

CrossLang

**Tom
Vanallemeersch**

CrossLang

Franklin Rooseveltlaan 348/bus 8, 9000 Gent, Belgium

[\[first name\].\[last name\]@crosslang.com](mailto:[first name].[last name]@crosslang.com)

Abstract

A well-known challenge of machine translation (MT) is accurately translating domain-specific terminology. While various methods have been suggested to address this challenge, they all come with limitations and increase the user's dependence on a specific MT engine. Recently, large language models (LLMs) for various natural language processing tasks, including automated translation, have gained significant attention, urging the need to investigate the potential of these models for terminology translation. Therefore, we compare ChatGPT, an LLM-based chatbot conversing with a user, to DeepL, an MT system converting sequences to sequences. We use both systems to perform translations with and without glossaries. We also combine both systems by post-editing MT output with the chatbot. Automated and manual evaluations indicate that the global translation quality of MT is better than or on par with that of the chatbot with a glossary, but that the latter system excels in terms of terminological accuracy when used for translation or for post-editing. While such post-editing avoids user dependence on a specific MT engine, it sometimes causes new translation issues, such as shifts in meaning, suggesting the need for future improvements. Our experiments focus on two language pairs, English-Russian and English-French, and on two domains (COVID-19 and legal documents).

1 Introduction

One of the most persistent and complex challenges in machine translation (MT) is the accurate handling of domain-specific terminology. This terminology is often context-specific, making its translation intricate and time-consuming. This problem has been approached with various MT techniques striving to improve the accuracy of translated terms, by enforcing a specific training procedure. However, each of these strategies faces limitations.

Lately, there has been a growing trend towards training large language models (LLMs) for specific tasks. For instance, GPT (Generative Pre-trained Transformer), when specialised for a

chatbot, can also be used to request a translation. Despite this trend, the potential of LLMs for terminology-aware translation remains relatively unexplored.

In this paper, we analyse the capabilities of a chatbot (more specifically ChatGPT 4), a system that *converses* (interacts) with a user, in handling the terminology problem. We compare it with an MT system (more specifically DeepL), which *converts* sequences to sequences. We also combine both systems, through a post-editing procedure. Our experiments involve two language pairs, English-Russian and English-French, and two domains, COVID-19 and legal documents.

The performance of the systems is evaluated using four scenarios. In the first one, an MT system is presented with English sentences containing challenging terms. In the second one, a chatbot is requested (prompted) to provide translations for these English sentences. In the third scenario, we provide the chatbot with a list of terms and their corresponding translations. The last scenario, post-editing, involves providing the chatbot with source sentences, MT output, and a glossary.

We assess the outcome of the tasks both automatically and manually. In the first case, we measure the global translation quality of a sentence using automatic metrics. In the second, we focus on terminological accuracy by assigning a sentence to an error category, where applicable.

In the subsequent sections, we describe the background of our research, the methodology, the data, and the results. Finally, we describe potential future workflow based on our findings and the challenges encountered.

2 Background

Several methods exist for incorporating terminology into neural machine translation (NMT) systems:

Mixing training data. This approach ensures the NMT training data contains both generic and domain-specific training data. While this allows the NMT engine to produce relevant terminology (e.g. to translate *bankruptcy* with French *faillite*), there is no guarantee of getting the right translation (e.g. *bankruptcy* may also be translated with French *banqueroute*).

Incorporating placeholders. This approach makes use of non-terminal tokens in NMT systems (such as $\langle term\#1 \rangle$), through pre- and postprocessing (Crego et al., 2016). While the NMT model learns how to deal with terminology, information is lost: the tokens constituting terms are no longer present during training of the model and thus inflection of target terms is not handled (Michon et al., 2020); this requires specific procedures before and after applying the NMT model.

Constrained decoding. This approach ensures that the desired term translations appear in the NMT output (Hokamp and Liu, 2017). This is achieved at the cost of higher computational demands, which slows down the translation process.

Injecting translations in the source sentence. The NMT model learns how to incorporate terminology translations in the target sentence when they are provided inside the source sentence (Song et al., 2019; Dinu et al., 2019). The system learns to copy words from time to time. This approach lacks the power to generalize, as the injection of the target term takes place without regard for the target context.

Recently, the use of LLMs for various natural language processing (NLP) tasks, including automated translation, has gained significant attention, urging the investigation of the potential of these models for terminology translation, especially in light of the above-mentioned drawbacks and the dependence of a user on a specific engine in terms of training data composition and training procedure.

NMT models and LLMs differ fundamentally in their training and architecture. NMT models are trained on parallel data and have an encoder-decoder architecture. In contrast, LLMs are trained on large amounts of monolingual data in one or more languages and employ a decoder-only architecture. This approach has opened new possibilities in multilingual NLP. LLMs are versatile, capable of being adapted for various tasks. For instance, they constitute the basis for chatbots such as ChatGPT, developed by OpenAI (Ouyang et al., 2022). Chatbots take user prompts and provide a response. Such prompts may also include labelled examples, which allows for in-context learning by the system (Brown et al., 2020).

The multilingual capabilities of LLMs are being investigated in comparison to NMT. For instance, Hendy et al. (2023) found that GPT models are very performant for high resource languages but have limited capabilities for low resource languages. In addition, Garcia et al. (2023) show the usefulness of providing a limited number of example sentence pairs when using an LLM to translate.

Various translation prompts have been proposed (Jiao et al., 2023). These prompts can either include translation task information only, provide additional context domain information, or use part-of-speech tags as auxiliary information (Gao et al., 2023).

Recent developments in NMT, such as those explored by Moslem et al. (2023), have begun addressing the integration of terminology in automated translation. This study notably improves the incorporation of pre-approved terms in translations using a methodology that combines synthetic data generation and terminology-constrained post-editing with LLMs like ChatGPT. However, the accurate rendering of terminology during translation by chatbots and comparison with traditional NMT systems at this level was not the primary focus of these studies. Our research, on the other hand, is specifically aimed at examining the efficacy of a chatbot as compared to an NMT system. We investigate how chatbots, which can use terminological information as context in prompts, without requiring specialised training data compositions or procedures, may not only facilitate terminology handling but also potentially enhance the quality of translation, especially in terms of terminological accuracy. The following sections will describe our approach in conducting this comparative analysis, highlighting the distinct aspects of our methodology.

3 Methodology

We compare the output of an NMT system (DeepL) to that of a chatbot (ChatGPT 4). We restrict the scope of our investigation to these two state-of-the-art systems, leaving the investigation of other systems (for instance open-source software) for future investigation. The NMT system and the chatbot are compared as follows. Based on a translation memory or parallel corpus in a specific domain, we select challenging source terms, select an illustrative sample of source sentences containing these terms, and apply the following scenarios to these sentences:

1. Translate the sentences using NMT.

2. Request the chatbot to translate the sentences. Prompts with the structure shown in Figure 11 are entered (we include multiple sentences in the prompt to provide more context).

3. Apply the same procedure as in 2 but include a glossary in the prompt, as shown in Figure 12. This is a form of in-context learning, as opposed to the zero-shot learning in scenario 2.

4. Given the sentences which, based on the human evaluation procedure described below, are known to be translated incorrectly at the terminological level by the NMT system, provide a prompt to the chatbot requesting it to post-edit the NMT output based on the glossary, as shown in Figure 13. The prompt also includes the source sentence.

Translate these sentences from English into Russian:

U.S. older adults, including those aged ≥ 65 years and particularly those aged ≥ 85 years, also appear to be at higher risk for severe COVID-19-associated outcomes; however, data describing underlying health conditions among U.S. COVID-19 patients have not yet been reported.

In the EU/EEA, the first three confirmed cases were reported by France on 24 January 2020 in persons returning from Wuhan, Hubei Province, China.

Figure 11. Prompt for scenario 2 (translation using chatbot)

Here is the list of COVID-19-related terms in English and their equivalents in Russian:

Respiratory distress syndrome – Острый респираторный дистресс-синдром

Respiratory dysfunction – дыхательная дисфункция

Given this glossary, translate the sentences provided below from English into Russian. Make sure that the terminology translation fully adheres to the glossary I provided, the translation domain is Covid-19.

Figure 12. Prompt for scenario 3 (translation using chatbot + glossary)

Please post-edit the following Russian sentences translated from English. The English source text is provided for your reference. The sentences are related to COVID-19, and I have noticed that the terminology used in the translations may not be accurate. Your task is to edit the sentences, replacing any incorrect or inadequate medical terms with the appropriate ones. You can refer to the list of COVID-19 terms provided below for guidance.

----- List of COVID-19 Terms (English to Russian) -----

...

English source:

...

Russian translation:

...

Please review each sentence carefully and make any necessary changes to ensure accurate and appropriate COVID-19 terminology is used.

Figure 13. Prompt for scenario 4 (post-edition of NMT output using chatbot + glossary)

We perform an automated evaluation of the global translation quality of the first three scenarios using the metrics BLEU, chr_f, TER and BERTScore. The first two of these calculate an n-gram match between output and reference (in terms of tokens or characters). The third one calculates post-editing effort and the fourth performs a semantic comparison of sentences using deep learning (embeddings).

We manually evaluate the translation output for all four scenarios (i.e. also the scenario for post-editing) at the terminological level, assigning one of the following error types to a sentence if applicable:

Inaccurate translation: the translated term (i) does not precisely match the original term's meaning despite maintaining the general sense, (ii) is misleading, or (iii) is unrelated to the source term.

Literal translation: the translated term matches the original term’s meaning but has a different, unusual phrasing.

Loss of elements: the system omits vital components of the source term in the translation.

4 Data

We apply the above methodology to two domains: COVID-19 (English-Russian) and legal-domain terminology (English-Russian and English-French).

For the first domain, we select a translation memory (TM) from the TICO-19 repository¹ and identify 49 challenging (that is, ambiguous, idiomatic, or culturally specific) terms in the English source sentences via SketchEngine.² We extract 90 sentences containing on average one or two of these terms from the TM, along with their Russian translations, to serve as reference translations.

For the second domain, we select 27 terms from the Rules of Court of the European Court of Human Rights, as well as from the European Convention on Human Rights, for both the English-Russian and English-French translation directions. In this case, we only perform (1) translation with the chatbot providing the glossary and (2) post-editing. We omit the use of the chatbot without the glossary, as the findings for the first domain, described in Section 4, clearly indicate a lower performance when working without a glossary.

4 Results

5.1 Automated Metrics Analysis

Table 26 shows the automated metric scores for the translation task involving COVID-19 terminology.

EN-RU COVID-19 (90 sentences)					
Metric	NMT	Chatbot	Delta with respect to NMT	Chatbot + glossary	Delta with respect to NMT
BLEU ↑	32.9	24.9	-8	29.0	-3.9
chr_f ↑	60.3	53.9	-6.4	59.8	-0.5
TER ↓	56.0	64.2	+8.2	59.0	+3
BERTScore ↑	88.2	86.1	-2.1	88.0	-0.2

¹ <https://tico-19.github.io>

² <https://www.sketchengine.eu>

Table 26. Automated metric scores for COVID-19

The results show that the NMT system outperforms the chatbot when the latter is not provided with a glossary, but that the gap shrinks when the glossary is provided. This demonstrates the efficacy of added contextual support.

Interestingly, in the legal domain (for which the chatbot is not tested without a glossary, as mentioned earlier), the chatbot with a glossary not only closes the gap with NMT but slightly outperforms it across all the evaluation metrics for both English-Russian and English-French translations (see Table 27). This indicates a pronounced effectiveness of the chatbot in handling domain-specific terminology when provided with a glossary.

(57 sentences)	EN-RU legal			EN-FR legal		
	NMT	Chatbot + glossary	Delta	NMT	Chatbot + glossary	Delta
BLEU ↑	33.3	33.8	+0.5	45.2	45.6	+0.4
chr_f ↑	61.2	62.0	+0.8	69.8	71.0	+1.2
TER ↓	55.4	55.0	-0.4	41.6	40.7	-0.9
BERTScore ↑	87.1	87.8	+0.7	89.6	90.5	+0.9

Table 27. Automated metric scores for the legal domain

5.2 Human Evaluation Results

The human evaluation focuses on assessing the accuracy of terminology in the translations.³ For the COVID-19 dataset, the use of the chatbot with a glossary markedly reduces terminological errors compared to NMT and chatbot without a glossary, as can be seen in Table 28.

EN-RU COVID-19 (90 sentences)					
Translation error type	NMT	Chatbot	Delta with respect to NMT	Chatbot + glossary	Delta with respect to NMT
Literal translation	16 (18%)	23 (26%)	+7 (8%)	0	-16 (18%)
Inaccurate translation	30 (33%)	46 (51%)	+16 (18%)	5 (5%)	-25 (28%)

³ The evaluation was performed by the authors. Potential future improvements consist of interannotator agreement and the involvement of domain experts.

Loss elements of	4 (4%)	5 (5%)	+1 (1%)	5 (5%)	+1 (1%)
All	50 (55%)	74 (82%)	+24 (27%)	10 (11%)	-40 (44%)

Table 28. Distribution of most common terminological error types in translations, COVID-19

Inaccurate translations drop dramatically from 33% with NMT and 51% with the chatbot without a glossary to just 5% with the chatbot plus a glossary. In the example below, the chatbot without a glossary translates *shortness of breath* as *одышка* instead of *затруднение дыхания* (the first Russian translation being a less severe term implying temporary breathlessness). The chatbot, when provided with the term list, provides the correct translation:

Source text: *Common symptoms include fever, cough and **shortness of breath**.*

Glossary: *shortness of breath* → *затрудненное дыхание*

NMT output: *Общие симптомы включают лихорадку, кашель и **одышку**. (literally: Common symptoms include fever, cough and **dyspnea**.)*

Translation by chatbot without glossary: *Распространенные симптомы включают лихорадку, кашель и **одышку**. (literally: Widespread symptoms include fever, cough and **dyspnea**.)*

Translation by chatbot with glossary: *Общие симптомы включают лихорадку, кашель и **затруднение дыхания**. (literally: Common symptoms include fever, cough and **shortness of breath**.)*

Similarly, literal translation errors are reduced to 0% with the chatbot plus a glossary, from 18% with NMT and 26% with the chatbot without a glossary. For instance, when translating the sentence provided below, the chatbot with a glossary chooses the more contextually appropriate term *самоизоляция*:

Source text: ... *try to stay indoors for **self-quarantine** and limit contact with potentially infected individuals.*

Glossary: *self-quarantine* → *самоизоляция*

NMT output: ... *помещениях для **самокарантина** и ограничить контакты с потенциально инфицированными людьми. (literal translation: term not commonly used)*

Translation by chatbot with glossary: ... *помещении для **самоизоляции** и ограничивать контакт с потенциально инфицированными лицами. (more appropriate term in this context)*

A similar pattern can be observed in the legal domain, where the chatbot with a glossary outperforms NMT in terms of terminological correctness (see Table 29), especially for English-French: errors of all types present in the NMT output are absent from chatbot output. Examples of sentences in the legal domain for both language pairs and their NMT and chatbot output are shown in Appendix A.

(57 sentences)	EN-RU legal			EN-FR legal		
	Scenario 1: NMT	Scenario 3: chatbot+glossary	Delta	Scenario 1: NMT	Scenario 3: chatbot+glossary	Delta
Literal translation	3 (5%)	4 (7%)	+1 (2%)	16 (28%)	0	-16 (28%)
Inaccurate translation	21 (37%)	16 (28%)	-5 (9%)	5 (9%)	0	-5 (9%)
Loss of elements	1 (2%)	2 (4%)	+1 (2%)	0	0	0
Total	25 (44%)	22 (39%)	-3 (5%)	21 (37%)	0	-21 (37%)

Table 29. Distribution of most common error types in translations, legal domain

5.3 Evaluation of Post-editing Task

We select all problematic NMT translations (50 sentences for the COVID-19 domain; 25 and 21 sentences for the legal texts for the English-Russian and English-French translation directions, respectively) and prompt the chatbot to post-edit them using the prompt specified in Figure 13. Table 30 shows the results for the post-editing task in both domains.

	EN-RU COVID-19	EN-RU legal	EN-FR legal
Sentences in NMT output containing error	50 (55% of 90)	25 (44% of 57)	21 (37% of 57)
Sentences with desired correction by chatbot	→ 47 (94%)	→ 15 (60%)	→ 20 (95%)

Table 30. Post-editing task results for COVID-19 and legal domains

For COVID-19, the chatbot successfully post-edits the terms in 47 sentences. An exemplary case of successful post-editing in this domain involves the term *transmissibility* in the following sentence:

Source text: *On one hand, the **transmissibility** of SARS-CoV-2 is at least as high as that of community-acquired HCoVs.*

Glossary: *transmissibility* → *передаваемость*

NMT output: *С одной стороны, **трансмиссивность** SARS-CoV-2, по крайней мере, так же высока, как и у HCoV, передающихся через сообщества. (literally: On one hand, the **transferability** of SARS-CoV-2 is at least as high as that of community-acquired HCoVs.)*

Translation by chatbot with glossary: *С одной стороны, **передаваемость** SARS-CoV-2, по крайней мере, так же высока, как и у HCoV, передающихся через сообщества.* (literally: *On one hand, the **transmissibility** of SARS-CoV-2 is at least as high as that of community-acquired HCoVs.*)

The NMT system translates the term as *трансмиссивность*, which is then correctly post-edited by the chatbot to *передаваемость*, adhering to the glossary's guidance.

The chatbot is able to handle inflections appropriately, ensuring that the translations are not only terminologically accurate but also grammatically coherent. In the example below, the chatbot not only replaces the term with the more accurate term *вспомогательных белков* from the glossary but also correctly adjusts the inflection to match the plural form used in the sentence:

Source text: *A number of lineage-specific **accessory proteins** are also encoded by different lineages of CoVs.*

Glossary: *accessory protein* → *вспомогательный белок*

NMT output: *Различные линии CoVs также кодируют ряд **специфических белков-аксессуаров**.* (literally: *Different lines of CoVs also code a number of specific **proteins-accessories**.*)

Post-edited with chatbot: *Различные линии CoVs также кодируют ряд **специфических вспомогательных белков**.* (literally: *Different lines of CoVs also code a number of specific **accessory proteins**.*)

The errors in the three incorrectly post-edited sentences involve an unchanged term, terms where information is lost, and a term where information is added. Apart from changes at the terminological level, the chatbot occasionally has agreement errors, loses important sentence elements, or changes sentence meaning, as in the following text, where it mistranslates *HCoV* as *hepatitis C*:

Source text: *It is also of particularly great interest to see whether SARS-CoV-2 might exhibit seasonality as in the cases of community-acquired **HCoVs**.*

Post-edited with chatbot: *Особый интерес представляет также вопрос о том, может ли SARS-CoV-2 проявлять сезонность, как в случае с внебольничным вирусом **genatuma C**.* (literally: *It is also of particularly great interest to see whether SARS-CoV-2 might exhibit seasonality as in the cases of community-acquired **hepatitis C**.*)

Moreover, as can be seen from Table 30, the success rate of post-editing varies across domains. More specifically, the lowest success rate is observed in the English-Russian legal domain, where the chatbot successfully post-edits 60% of the sentences. In contrast, the chatbot achieves its highest success rate in the English-French legal domain, successfully correcting errors in 95% of the sentences, with only one error remaining.

Appendix A lists automatically post-edited sentences in the legal domain for both language pairs. Appendix B shows automatic metrics and confidence intervals for the various types of output (NMT, chatbot with glossary, chatbot post-editing) for this domain and these language pairs.

6 Conclusions and Future Work

In order to improve the performance of NMT engines in the area of term translation, various strategies have been developed for training data composition and training setup. As these requirements can lead to a dependence of the user on a specific engine, we explored to what extent the multilingual capabilities of a chatbot, a system which can be provided with various prompts in a user-friendly way, are useful for improving terminological accuracy. To this end, we compared the output of DeepL, an NMT engine, to that of a chatbot, ChatGPT.

Our comparative analysis was structured through four distinct scenarios, designed to evaluate translation outputs in terms of global translation quality and terminological accuracy: (i) translation using NMT without any additional input or modification; (ii) translation by the chatbot without the aid of a glossary; (iii) translation by the chatbot with the help of a glossary; (iv) post-editing of NMT output using the chatbot. To assess the performance across these scenarios, we employed both automated evaluation metrics and human evaluation methods.

On the one hand, NMT offers better translation quality (in the COVID-19 domain for English-Russian) or on-par output (for the legal domain in English-Russian and English-French) compared to the chatbot with a glossary. On the other hand, the latter system excels in terminological accuracy when requested to translate or to post-edit NMT output; this is especially the case for the COVID-19 domain and for the English-French legal-domain text. However, post-editing also carries the risk of introducing noise.

Our findings suggest opportunities for further research and development:

To reduce the risk of noise introduction, we could vary prompt phrasing, the number of sentences included in a prompt, and the type and size of context. For instance, we may vary the number of glossary terms and include example sentence pairs for terms with multiple translations in a domain.

Given the variety of NMT systems and chatbots (variants of ChatGPT, open-source models, etc.) and the fast-paced evolution in the area of chatbots, we could extend the scope of the study by increasing the number of systems, as well as the number of domains, languages, and metrics (e.g. additional deep learning metrics). It is also important to investigate whether and when “hallucinations” (nonsensical output) appear, which are regularly observed in complex deep learning based systems.

While our study aimed to reduce dependence on specific MT engines, it would be interesting to compare the performance of a chatbot-based approach to that of uploading a user-specified glossary in a commercial system like DeepL.

Acknowledgements

This work is the result of an internship of the main author at CrossLang in the first half of 2023. She would like to express her gratitude for having received the opportunity to investigate the topic described in this paper.

References

- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1877–1901.
- Crego, Josep, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran’s Pure Neural Machine Translation Systems. arXiv:1610.05540.
- Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training Neural Machine Translation to Apply Terminology Constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068.
- Gao Yuan, Ruili Wang, and Feng Hou. 2023. How to Design Translation Prompts for ChatGPT: An Empirical Study. arXiv:2304.02182.
- Garcia, Xavier, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The Unreasonable Effectiveness of Few-shot Learning for Machine Translation. In *Proceedings of the International Conference on Machine Learning*, pages 10867–10878.
- Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good are GPT models at Machine Translation? A Comprehensive Evaluation. arXiv:2302.09210.
- Hokamp, Chris, and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation using Grid Beam Search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546.
- Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? Yes with GPT-4 as the engine. arXiv:2301.08745.
- Michon, Elise, Josep Crego, and Jean Senellart. 2020. Integrating Domain Terminology into Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937.
- Moslem, Yasmin, Gianfranco Romani, Mahdi Molaei, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. arXiv:2310.14451.
- Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of Advances in Neural Information Processing Systems*, pages 27730–27744.
- Song, Kai, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for Enhancing NMT with Pre-specified Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459.

Appendix A: Examples for EN-RU and EN-FR (Legal Documents)

EN-RU :

Glossary	Source	Chatbot translation	NMT	Chatbot post-editing
Removal – Выдворение	The Court must be notified of any change to those details (date and time of removal , address etc.) as soon as possible.	Суд должен быть немедленно уведомлен обо всех изменениях этих данных (дата и время выдворения , адрес и т.д.). (literally: The Court must be notified of any change to those details (date and time of removal , address etc.) as soon as possible.)	Суд должен быть уведомлен о любых изменениях в этих данных (дата и время удаления , адрес и т.д.) в кратчайшие сроки. (literally: The Court must be notified of any change to those details (date and time of deletion , address etc.) as soon as possible.)	Суд должен быть уведомлен о любых изменениях в этих данных (дата и время выдворения , адрес и т.д.) как можно скорее. (literally: The Court must be notified of any change to those details (date and time of removal , address etc.) as soon as possible.)
Amendment – Поправка	Any proposal for amendment shall be communicated by the Secretary General of the Council of Europe to the Parties.	Любое предложение о поправке должно быть сообщено Генеральным секретарем Совета Европы Сторонам. (literally: Any proposal for amendment shall be communicated by the Secretary	Любое предложение о внесении изменений доводится Генеральным секретарем Совета Европы до сведения Сторон. (literally: Any proposal for change shall be communicated by the Secretary	Любое предложение о Поправке доводится Генеральным секретарем Совета Европы до сведения Сторон. (literally: Any proposal for Amendment shall be communicated by the Secretary

		General of the Council of Europe to the Parties.)	General of the Council of Europe to the Parties.)	General of the Council of Europe to the Parties.)
Judgment – постановление	The final judgment shall be published.	Окончательное суждение будет опубликовано. (literally: The final opinion shall be published.)	Окончательное решение должно быть опубликовано. (literally: The final decision shall be published.)	Окончательно постановление должно быть опубликовано. (literally: The final judgment shall be published.)
Application – постановление	The Court shall reject any application which it considers inadmissible under this Article.	Суд отклонит любую жалобу , которую он считает недопустимой в соответствии с этой статьёй. (literally: The Court shall reject any complaint which it considers inadmissible under this article.)	Суд отклоняет любое заявление , которое он считает неприемлемым в соответствии с настоящей статьёй. (literally: The Court shall reject any application which it considers inadmissible under this article.)	Суд отклоняет любое заявление , которое он считает неприемлемым в соответствии с настоящей статьёй. (literally: The Court shall reject any application which it considers inadmissible under this article.)

Glossary (continued)	Source	Chatbot translation	NMT	Chatbot post-editing
Domestic court – внутригосударственный Суд	The Court does not hear appeals against decisions of domestic courts , and	Суд не рассматривает апелляции по решениям национальных судов , и	Суд не рассматривает апелляции на решения национальных судов , и	Суд не рассматривает апелляции на решения национальных судов , и

	<p>applicants in expulsion or extradition cases should pursue domestic remedies [...]</p>	<p>заявители в делах об исключении или экстрадиции должны искать национальные средства правовой защиты [...]</p> <p>(literally: The Court does not hear appeals against decisions of national courts, and applicants in exclusion or extradition cases should pursue national means of legal protection [...])</p>	<p>заявители по делам о высылке или экстрадиции должны использовать внутренние средства правовой защиты [...]</p> <p>(literally: The Court does not hear appeals against decisions of national courts, and applicants in expulsion or extradition cases should pursue internal means of legal protection [...])</p>	<p>заявители по делам о высылке иностранцев или экстрадиции должны использовать Внутригосударственные Средства Защиты [...]</p> <p>(literally: The Court does not hear appeals against decisions of national courts, and applicants in expulsion or extradition cases should pursue domestic means of legal protection [...])</p>
--	--	---	--	--

EN-FR :

Glossary	Source	Chatbot translation	NMT	Chatbot post-editing
Applicant – réquerant	In all cases before a Chamber or the Grand Chamber, a High Contracting Party one of whose nationals is an applicant shall have the right to submit	Dans tous les cas devant une Chambre ou la Grande Chambre, une Haute Partie Contractante dont un national est réquerant a le droit de soumettre des observations écrites et de	Dans toutes les affaires portées devant une chambre ou la Grande Chambre, une Haute Partie contractante dont l'un des ressortissants est demandeur a le droit de présenter des	Dans toutes les affaires devant une chambre ou la Grande Chambre, une Haute Partie contractante dont l'un des requérants est ressortissant a le droit de soumettre des observations écrites et de

	<p>written comments and to take part in hearings.</p>	<p>participer aux audiences.</p> <p>(literally: In all cases before a Chamber or the Grand Chamber, a High Contracting Party of which a national is an applicant has the right to submit written comments and to take part in hearings.)</p>	<p>observations écrites et de participer aux audiences.</p> <p>(literally : In all cases brought before a Chamber or the Grand Chamber, a High Contracting Party of which one of the nationals is applicant has the right to present written comments and to take part in hearings.)</p>	<p>participer aux audiences.</p> <p>(literally : In all cases before a Chamber or the Grand Chamber, a High Contracting Party of which one of the applicants is national has the right to submit written comments and to take part in hearings.)</p>
--	---	---	---	---

Glossary (continued)	Source	Chatbot translation	NMT	Chatbot post- editing
Removal – renvoi	The Court must be notified of any change to those details (date and time of removal , address etc.) as soon as possible.	Le tribunal doit être informé de tout changement de ces détails (date et heure de renvoi , adresse, etc.) dès que possible. (literally : The Court must be informed of any change to those details (date and hour of removal , address etc.) as soon as possible.)	Toute modification de ces données (date et heure du déménagement , adresse, etc.) doit être notifiée à la Cour dans les plus brefs délais. (literally : Each modification of these data (date and hour of the relocation , address etc.) must be notified to the Court in the shortest delays.)	Toute modification de ces informations (date et heure du renvoi , adresse, etc.) doit être notifiée à la Cour dans les plus brefs délais. (literally : Each modification of this information (date and hour of the removal , address etc.) must be notified to the Court in the shortest delays.)
Lawful detention – détention régulière	the lawful detention of persons for the prevention of the spreading of infectious diseases, of persons of unsound mind, alcoholics or drug addicts or vagrants;	la détention régulière de personnes pour la prévention de la propagation de maladies infectieuses, de personnes d'esprit dérangé, d'alcooliques ou de toxicomanes ou de vagabonds; (literally : the lawful detention of persons for the prevention of the spreading of	la détention légale de personnes en vue de prévenir la propagation de maladies infectieuses, d'aliénés, d'alcooliques, de toxicomanes ou de vagabonds ; (literally : the lawful detention of persons in order to prevent the spreading of infectious diseases, of	la détention régulière de personnes pour prévenir la propagation de maladies infectieuses, de personnes d'esprit dérangé, d'alcooliques, de toxicomanes ou de vagabonds ; (literally : the lawful detention of persons for preventing the spreading of infectious diseases, of

		infectious diseases, of persons of disturbed mind, of alcoholics or of drug addicts or of vagrants;)	lunatics, of alcoholics, of drug addicts or of vagrants;)	persons of disturbed mind, of alcoholics, of drug addicts or of vagrants;)
Written pleadings – observations écrites	Rule 38 – Written pleadings	Règle 38 – Observations écrites (literally : Rule 38 – Written comments)	Article 38 – Plaidoiries écrites (literally : Article 38 – Written pleadings)	Règle 38 – Observations écrites (literally : Rule 38 – Written comments)
Expenditure on the court – frais de fonctionnement de la Cour	Article 50 – Expenditure on the Court The expenditure on the Court shall be borne by the Council of Europe.	Article 50 – Frais de fonctionnement de la Cour Les frais de fonctionnement de la Cour sont à la charge du Conseil de l'Europe. (literally : Article 50 – Costs for functioning of the Court The costs for functioning of the Court should be borne by the Council of Europe.)	Article 50 – Dépenses de la Cour Les dépenses de la Cour sont supportées par le Conseil de l'Europe. (literally : Article 50 – Expenses of the Court The expenses of the Court are supported by the Council of Europe.)	Article 50 – Frais de fonctionnement de la Cour Les frais de fonctionnement de la Cour sont supportés par le Conseil de l'Europe. (literally : Article 50 – Costs for functioning of the Court The costs for functioning of the Court are supported by the Council of Europe.)

Appendix B: Automatic Metrics for EN-RU and EN-FR (Legal Documents)

EN-RU:

System	bertscore ($\mu \pm 95\%$ CI)	BLEU ($\mu \pm 95\%$ CI)	chrF2 ($\mu \pm 95\%$ CI)	TER ($\mu \pm 95\%$ CI)
NMT	87.3 (87.3 \pm 1.5)	33.6 (33.8 \pm 5.7)	61.4 (61.5 \pm 3.3)	54.9 (54.7 \pm 6.1)
Chatbot translation with glossary	85.3 (85.3 \pm 1.3) (p = 0.0020)*	25.8 (25.9 \pm 4.5) (p = 0.0010)*	55.4 (55.4 \pm 3.1) (p = 0.0010)*	63.2 (63.1 \pm 5.5) (p = 0.0010)*
Chatbot post-editing	87.8 (87.8 \pm 1.4) (p = 0.1179)	33.8 (34.0 \pm 5.9) (p = 0.3117)	62.0 (62.1 \pm 3.0) (p = 0.1988)	55.0 (54.8 \pm 5.9) (p = 0.3796)

EN-FR:

System	bertscore ($\mu \pm 95\%$ CI)	BLEU ($\mu \pm 95\%$ CI)	chrF2 ($\mu \pm 95\%$ CI)	TER ($\mu \pm 95\%$ CI)
NMT	89.6 (89.6 \pm 1.7)	89.6 (89.6 \pm 1.7)	89.6 (89.6 \pm 1.7)	89.6 (89.6 \pm 1.7)
Chatbot translation with glossary	45.2 (45.1 \pm 5.3)	45.2 (45.1 \pm 5.3)	45.2 (45.1 \pm 5.3)	45.2 (45.1 \pm 5.3)
Chatbot post-editing	69.8 (69.8 \pm 3.4)	69.8 (69.8 \pm 3.4)	69.8 (69.8 \pm 3.4)	69.8 (69.8 \pm 3.4)

Hierarchical Data Linkage in a Terminology Management System: Challenges and Solutions at Bioleksipēdija

Karina Šķirmante

Ventspils University of Applied
Sciences

karina.krinkele@venta.lv

Silga Svīķe

Ventspils University of Applied
Sciences

silga.svike@venta.lv

Arturs Stalažs

Institute of Horticulture
arturs.stalazs@llu.lv

Gints Jasmonts

Ventspils University of Applied
Sciences

gints.jasmonts@venta.lv

Roberts Ervīns Ziediņš

Ventspils University of Applied
Sciences

s20ziedrobe@venta.lv

Abstract

Since 2021, a collaborative team of terminologists, translators, and information system developers have worked together to develop a new open-access, interactive, multifunctional, terminology management system Bioleksipēdija. The system is developed for special lexis data storage and a wide range of statistical and search options designed for language research purposes and comparative multilingual linguistic studies. The system developed is a terminology management tool, published in January 2024 under the domain bioleksipedija.lv.

This article describes the systematic taxa tree module of the system developed, with integrated hierarchical data linkage. This module links scientific or Latin names of organisms within a systematic tree structure while also incorporating vernacular names of organisms linked to publications. The module developed allows translators to search for and analyse precise terminology, considering both taxon placement within the systematic tree and its frequency of use, as measured by mentions in real publications. The system is used for data collection and on 19.08.2024, stored an overall, 74,954 scientific and 96,692 vernacular names of organisms, 1,780 names of diseases caused by organisms, 3,127 dictionary words, 429 terms and 645,582 linkages in 9,718 bibliography units.

1 Introduction

The authors of this paper compare various tools, including World Flora Online (<http://www.worldfloraonline.org/>), Latvian National Terminology Portal (<https://termini.gov.lv/>), Letonika.lv (<https://www.letonika.lv/>), Tēzaurus.lv (<https://tezaurus.lv/>), Latvijasdaba.lv (<https://www.latvijasdaba.lv/>), Periodika.lv (<http://periodika.lv/>), in their research (Stalažš et. al. 2023) as these are by far the most frequent

websites used by translators when translating texts containing the names of organisms. This statement is based on the personal experience of the authors, who are also practising translators. Although the research of words forms, terms and names of organisms is partly supported by the Latvian National Terminology Portal and Periodika.lv, the authors found that, only in the field of biology, there are no specific open-access collections of data that allow the checking of a large number of organism names for changes over time. This prompted the development of such a tool. To achieve this functionality, since 2021, a collaborative team of terminologists, translators, and information system developers have worked together to develop a new tool — Bioleksipēdija (Šķirmante et. al. 2024) which is open access, web-based, novel interactive multi-functional research database-management system with a wide range of statistical and search possibilities suitable for language research and public use worldwide.

The Bioleksipēdija includes organism names, cultivar names, dictionary entries, biology terms with definitions. These entries are tailored to specific publications and languages. An important feature of the tool is that it includes the scientific names of organisms linked within the systematic tree. This is an important option for translators who need to search for equivalents and to learn the exact correspondence of terms, in particular — plant or animal names, in the source language when translating a text from the target language. The Bioleksipēdija supports data entry in 184 languages, all of which adhere to the standardized ISO 639 nomenclature for languages. Any combination of languages is possible, but the hierarchy tree ensures the accuracy of the scientific name of the organism. As of August 2024, data entries have been stored in 26 languages. Figure 1 illustrates the complete entries, highlighting the 11 most common languages. Since the tool's development was funded by the Latvian Council of Science, the primary emphasis has been on research conducted in the Latvian language. This focus is clearly illustrated by the predominance of Latvian language data entries in Figure 1. Nevertheless, the tool remains fully applicable to research involving other languages recognized by the ISO 639 standard.

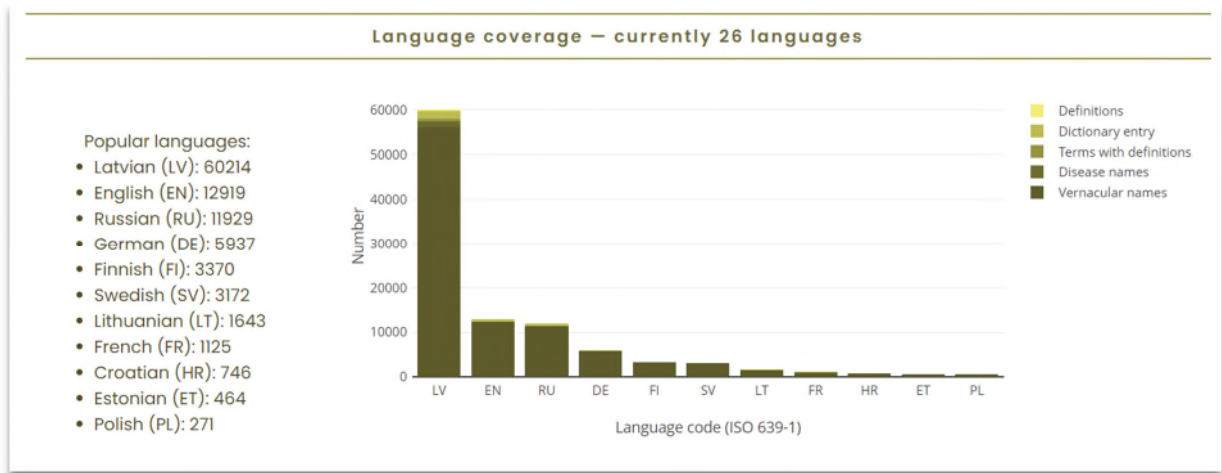


Figure 1. Languages covered and entries in Bioleksipēdija.

The Bioleksipēdija provides the opportunity not only to use the existing data collection, collected during the project including 9,718 bibliography units, 176,982 entries and 9,807 systematic tree entries (data as of 19.08.2024), but also to create new collections within specific

research topics. The Bioleksipēdija offers multiple search options — 1) multilingual search where Bioleksipēdija supports advanced search options, enabling researchers to retrieve specific entries across multiple languages, facilitating cross-linguistic comparisons, 2) search in the systematic tree of organisms, to better evaluate the name searched. The Bioleksipēdija not only allows the retrieval of data within the system but it also enables the export of data to an Excel file, which can be used for further research integrating with other tools. The Bioleksipēdija consists of twelve modules which have already been described (see Šķirmante et. al. 2024), so only one module — hierarchical data linkage module — will be described in detailed in this paper.

2 Hierarchical data linkage module

2.1 General information

The Bioleksipēdija offers the opportunity to explore names of organisms and their usage in various publications over time, currently searchable for the time period from 1924 to 2023. This type of data collection allows research on the frequency of use and variants of organism names, as each vernacular plant or animal name is linked to the scientific name, thus enabling the easy tracking of changes in frequency and variations during the exact time periods as well as statistics. Additionally, the Bioleksipēdija allows users to view organism names within the systematic tree, where data is organized hierarchically within a tree-like data structure. This enables users to retrieve scientific names and their linked vernacular names from the database collection.

To enhance the functionality of Bioleksipēdija, a hierarchical data linkage module was developed, serving as a middleware between two data collections —

1) data organized within a tree-like structure, facilitating a hierarchical representation. This collection establishes linkages between scientific names of organisms in various taxonomic rankings within a total of 34 taxonomic levels. The module consists of five distinct categories (kingdoms): plants, animals, fungi, bacteria, and viruses. An illustrative example featuring 'Quercus robur' demonstrating the visualization of the systematic tree in Bioleksipēdija is presented in Figure 2.

2) the data collection where the linkage between the organism name and bibliography unit is stored with additional information, for example, language mark, bibliography page number, user data, date and time, user comments. Refer to Figure 3 for an illustrative list of publications that include the usage of 'Quercus robur' within the scope of Bioleksipēdija.

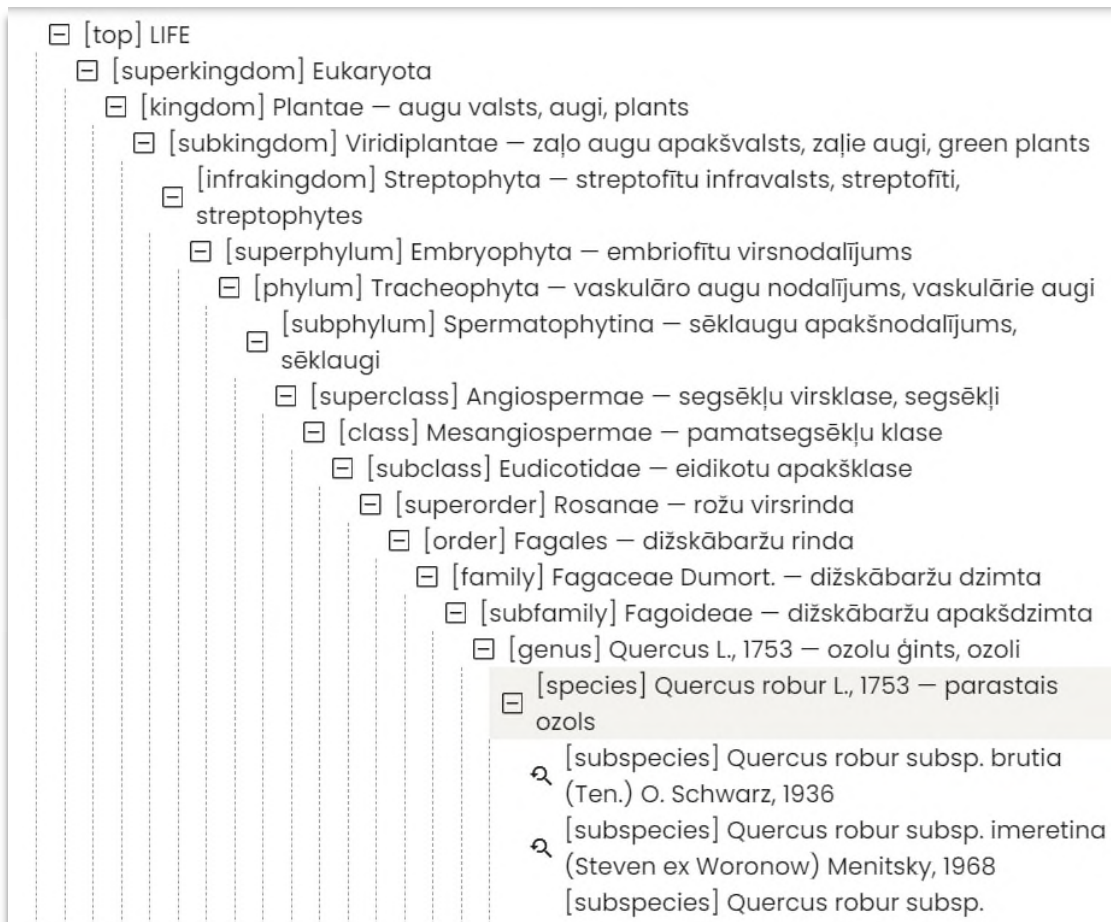


Figure 2. Systematic tree view of the searched '*Quercus robur*'.

Year	Bibliographic
1950	Galenieks, P., (1950). Botaniskā vārdnīca. Rīga: Latvijas Valsts izdevniecība, 218
1950	Vaņins, S., (1950). Koksnes zinātne. Rīga: Latvijas Valsts izdevniecība, 464
1956	Balode, K., Galenieks, F., (1956). Biškopība. Trešais pārstrādātais izdevums. Rīga: Latvijas Valsts izdevniecība, 475
1959	Barons, K., (1959). Koki un krūmi dārzos. Daba. Dārzs. Drava (5), 18–20
1960	Šmits, V., (1960). Mizgrauži. Latvijas PSR teritorijā sastopamās Ipidae sugas. Rīga: Latvijas Valsts izdevniecība, 208
1965	Rupais, A., (1965). Dārzkopībā kaitīgās laputis un to apkarošana. Rīga: Zinātne, 92
1965	Markovs, M., (1965). Vispārīgā ģeobotānika. Rīga: Liesma, 435
1970	Jaunpucis, S. (resp. ed.), (1970). Puķu, dekoratīvo augu un to iztīrījumu mazumtirdzniecības cenu cenrādis. Rīga: Latvijas PSR Komunālās saimniecības ministrija, 77
1973	Langenfelds, V., Ozoliņa, E., Ābele, G., (1973). Augstāko augu sistematika. Rīga: Zvaigzne, 406
1981	Laiviņa, S., Laiviņš, M., (1981). Grīņu rezervāta augu sabiedrību struktūra un vides faktori. Mežsaimniecība un Mežrūpniecība (3), 16–21
1982	Eniņš, G., (1982). Koks — dabas pieminekļi. Rīga: Zinātne, 95
1984	Laiviņš, M., (1984). Latvijas PSR ezeru salu baltalkšņu mežu sabiedrības. Mežsaimniecība un Mežrūpniecība (6), 23–27
1986	Spuris, Z. (ed.), (1986). Ceļvedis pa ZA Botānisko dārzu. Rīga: Zinātne, 117
1986	Zvirgzds, A., (1986). Koks pilsētā. Rīga: Zinātne, 96
1986	Hrzanovskis, V., Ponomarenko, S., (1986). Botānika. Rīga: Zvaigzne, 438
1988	Āķis, E. (comp.), (1988). Krāšņumaugu, augļu koku un ogulāju stādu sortiments Latvijas PSR kokaudzētāvās 1988. gadā. Rīga: Jūrmalas Zaļumsaimniecības eksperimentālās ražošanas kombināts, 57
1988	Kaškura, A., Šmaukstelis, E., (1988). Koku un krūmu pavairošana ar spraudņiem. Apskats. Rīga: LatZTIZPI, 54
1991	Cinovskis, R., Mauriņš, A., Zvirgzds, A., (1991). Skāiveru dendrārijs. Ceļvedis. Rīga: Zinātne, 117
1991	Liepa, I., Mauriņš, A., Vimba, E., (1991). Ekoloģija un dabas aizsardzība. Rīga: Zvaigzne, 301
1992	Boisset, C., (1992). The plant growth planner. 200 illustrated charts for shrubs, trees, climbers and perennials. London: Mitchell Beazley
1994	Kabucis, I., (1994). Baltijas ģeobotāniskā province. In: G. Kavacs (ch. ed.) Latvijas daba, enciklopēdija: Lettische Naturencyklopädie: The
1994	Laiviņš, M., (1994). Barkavas ozolu audze. In: G. Kavacs (ch. ed.) Latvijas daba, enciklopēdija: Lettische Naturencyklopädie: The

Figure 3. Part of the publication list of information about '*Quercus robur*' is collected and covered in Bioleksipēdija

The hierarchical data linkage module developed is integrated with the statistics module, thus enabling the comparison of data between both collections. This middleware module facilitates the retrieval of hierarchical statistics related to the scientific names of organisms. It provides access to all linked vernacular names of organisms, along with additional details such as the number of publications mentioning this linkage, the languages, and the publication's earliest and most recent years covered by Bioleksipēdija (see Figure 4). This functionality allows translators to analyse the historical usage patterns of a searched name over time.

Chosen taxon:
Quercus robur L., 1753 [species]
 Term last verified: 29.07.2023.

LV: parastais ozols

Valoda
 Anglu Latviešu

SEARCH STATISTICS

Title	Publication Count ↓	Min Year	Max Year
parastais ozols	68	1924	2023
ozols	16	1924	2021
pedunculate oak	4	1999	2022
oak	3	1997	2017
Pedunculate Oak	2	1997	2005
common oak	2	1998	2004
Pedunculate oak	2	2006	2007

Figure 4. Plant species '*Quercus robur*', linked to local names of organisms within publications covered by Bioleksipēdija and filtered by English and Latvian.

2.2 Overview of used technologies

The association of organism names with publications is a non-hierarchical linkage and is a standard relational association. Consequently, this information is stored in a relational-type database organized into tables. To achieve this requirement, the system uses a MySQL relational database. MySQL is one of the top database engines, as it is efficient and easy to use. It guarantees constant uptime, which is critical for a web-based system (Győrödi 2020). MySQL allows quick data access to extract statistics of names of organisms, but it is not efficient when dealing with hierarchical data. Organizing scientific names in a systematic tree requires a more effective storage method than that offered by typical relational databases. Therefore, different NoSQL database types were studied and evaluated to find a suitable solution. (Rai and Chettri 2018) emphasized that MongoDB is the preferred choice for hierarchical data storage. MongoDB is a source-available cross-platform document-oriented database program, which provides a solution for organizing data using a tree structure. As a result, MongoDB was chosen as the secondary database for Bioleksipēdija, where hierarchical linkages between scientific names of organisms are organized in documents within a tree-like structure. In the MongoDB database, scientific names, their spelling variants, and synonyms

are stored. This allows the extraction of recognised scientific names and also the review of spelling variants to assess the correct one. Spelling variants and synonyms are extracted from publications and linked to recognised scientific names. It is also possible to examine local names of organisms associated with the recognised name. To develop this functionality, a separate table in the MySQL database was created, where the MongoDB database's ID of the recognised scientific name of organism is linked with the MySQL database's ID of the associated local name of organism, which is also linked to publications. This provides statistics on how often a particular recognised name is used in publications alongside specific local names of organisms. The use of two different types of databases allows us to efficiently retrieve data queries. However, it also imposes the obligation to synchronize both databases to avoid situations where a name is deleted in one database but still exists in the other. In such a scenario, data retrieval in the statistics module could yield inaccurate results.

2.3 Challenges and Solutions

One of the challenges within the module developed is homonymy, where the same scientific name of organisms can be applied across multiple systematic trees. For example, '*Pieris*' may be both 1) a plant genus, linked with the vernacular name 'kalnērikas' in Latvian ('fetterbush or pieris' in English) and 2) an animal genus, linked with the vernacular name 'balteņi' in Latvian ('whites or garden whites' in English). To deal with this problem, a dedicated database segment equipped with a corresponding service layer was established for storing, deleting, editing, and extracting data.

The second challenge involved ensuring effective synchronization between both databases. Currently, this challenge is tackled thanks to the development of a Python script responsible for daily synchronization between the databases, managing scientific and vernacular names of organisms, and making necessary table adjustments. In the future, two-phase commit transactions could be implemented.

The third challenge is centred around optimizing data retrieval from both databases. This was accomplished by a Spring microservices architecture that we used in the system development. It allows use of each module as an independent microservice, and their communication is facilitated through distinct gateway project.

3 Short Summary

During this project, a new tool — terminology management system — Bioleksipēdija was developed. The design of the Bioleksipēdija was a challenging task to ensure effective performance in all operations due to the data linkages in multiple directions — organism names linkage to bibliography units, linkage to other synonymic names of organisms, and linkage in the systematic tree. The main challenge of the module for hierarchical data linkage was homonymy where the same scientific name of organism can be applied across multiple systematic trees, e.g., it can appear in two kingdoms — plants and animals. It is crucial for system users to filter only those vernacular names of organisms that belong to specific distinct categories (kingdoms), regardless of the homonymy of the name. Bioleksipēdija was developed for translators working on the translation of texts related to biology, to help them use the most appropriate and exact terminology.

Acknowledgements

This research has been funded by the Latvian Council of Science, project ‘Smart complex of information systems of specialized biology lexis for the research and preservation of linguistic diversity’, No. lzp-2020/1-0179.

References

- Győrödi, C., Dumșe-Burescu, D., Zmaranda, D. Győrödi, R., Gabor, G., Pecherle, G. (2020). *Performance Analysis of NoSQL and Relational Databases with CouchDB and MySQL for Application's Data Storage*, Applied Sciences, 10(23), 8524.
- Rai, R., Chettri, P. (2018). *Chapter Six - NoSQL Hands On*, Editor(s): Pethuru Raj, Ganesh Chandra Deka, Advances in Computers, Elsevier, Volume 109, Pages 157-277.
- Stalažš, A., Šķirmante, K., Sviķe, S., Jasmonts, G., Ziediņš, R. E. (2023). *Experience of design and development of a new open access interactive multifunctional database management system for special lexis of biology*, Studies About Languages, Issue 42, Pages 52-67.
- Šķirmante, K., Jasmonts, G., Ziediņš, R. E., Sviķe, S., Stalažš, A. (2024). *New Open Access Interactive Multifunctional Database Management System for Research of Biological Terminology: Technical Solutions*. SpringerLink, Advanced Research in Technologies, Information, Innovation and Sustainability, 1935 CCIS, Pages 282–296.

Exploring the integration of ChatGPT in academia and in the office: a preliminary case study

Kyriaki Kourouni

Aristotle University of Thessaloniki,

GR-54124 Thessaloniki, Greece

kkourouni@enl.auth.gr

Abstract

Recent technological developments are having a great impact in the translation world. Institutions offering translator training programmes are called upon to catch up by modifying their curricula accordingly, without, however, having a clear idea of how these developments will unfold in the near future. Freelancers and small translation companies are also urged to step up and adapt to the new, fluid, reality.

This paper forms part of a larger study on the status of AI in the Greek-speaking translation landscape. It reports on work-in-progress focusing on whether translator trainers as well as freelancers/small translation companies in Greece and Cyprus have started integrating ChatGPT into their programmes or workflows, respectively, and how, by means of questionnaires as a first step, to detect areas where both groups might collaborate toward paving a (more) rewarding future for everyone involved. Findings indicate that both groups have started testing ChatGPT, often with satisfactory results, and they have rather mixed views for the future in relation to ChatGPT and other Generative AI tools.

1 Introduction

The study on the status of AI in the Greek-speaking translation landscape was inspired by the ongoing technological leaps taking place during 2023, rapidly modifying both the translator training environment as well as the translation industry, from course syllabi to everyday working mode. Talks during meetings, for example within the framework of the European Master in Translation Working Group on Translation Tools and Technologies (2019-2024) and Translating Europe Forum 2023, as well as more locally, within Greece, during departmental meetings and events held by professional associations such as the Panhellenic Association of Translators have highlighted the need for a clearer picture.

The aim of this paper is to present, as accurately as possible, a mapping of the situation in Greece, starting with the attitude of translator trainers and translation professionals, as a basis for future proposals for satisfactory environments, based on data.

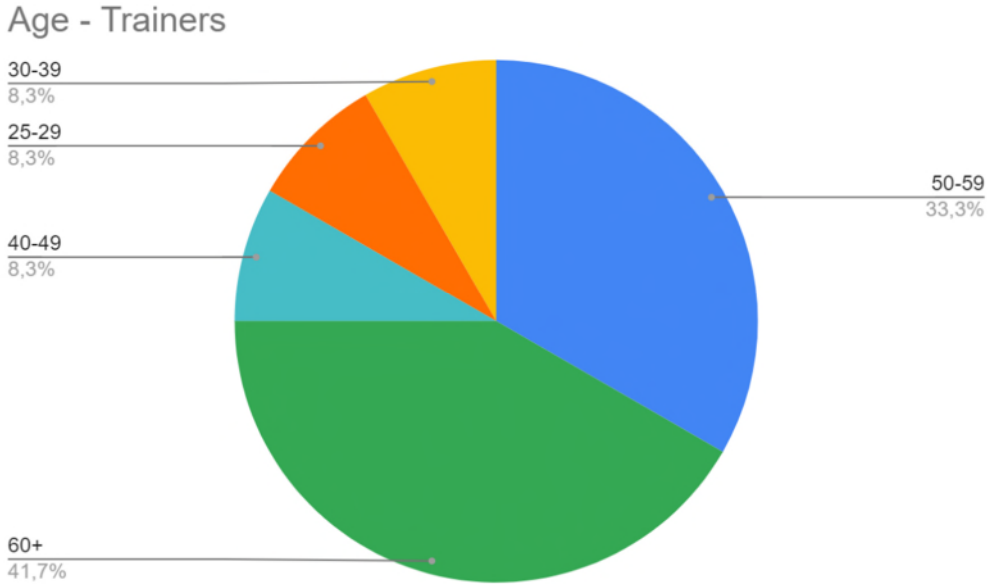
2 Methodology

The study focused on two groups, namely translator trainers and freelance translators/small translation companies, based in Greece and Cyprus. The list of translator trainers was compiled following searches on websites from state universities and private translation schools. The translator trainer group was invited to participate via email. Regarding translation professionals, a request was made to professional associations who, in turn, forwarded the message to their members. Another request was made via Facebook to relevant Greek translator groups.

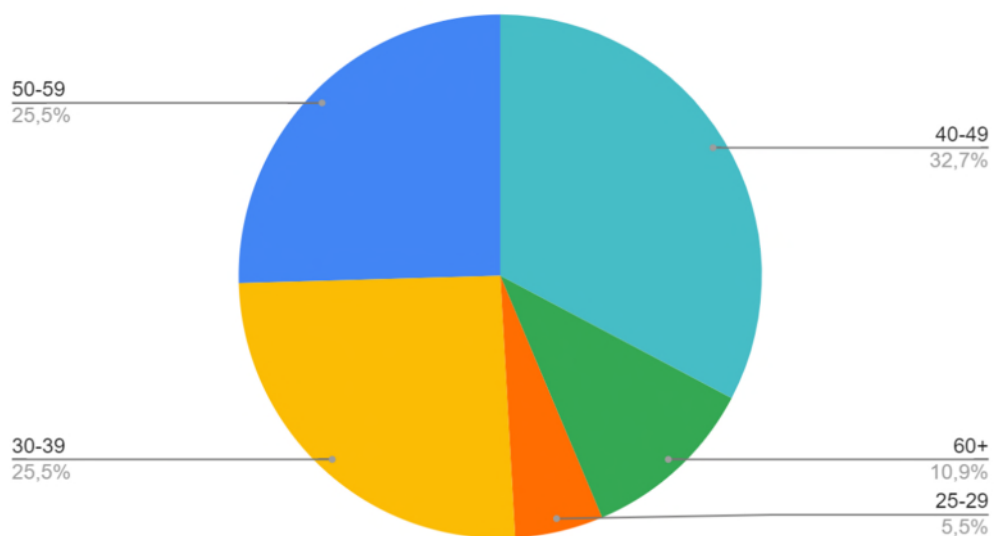
The study used an ad hoc blind questionnaire covering demographic variables as well as questions related to the use of and attitude toward ChatGPT. Questionnaire items covered use of and attitudes toward ChatGPT in general, rather than through a thorough examination of various perspectives of the platform features, to keep questionnaire completion within 10 minutes and thus, make the questionnaire more attractive to prospective respondents. The questionnaire form (one form per group albeit with identical items) was sent out in November 2023. The paper focuses on answers by members of both groups as to whether they have been using ChatGPT, for what purpose and/or task, and whether they are satisfied by the end result. The questionnaire also includes a few more general questions regarding their attitude in relation to AI. The results presented here derive from a sample of 12 trainers and 55 professionals, respectively.

2.1 Demographic data

Regarding age, 75% of translator trainers (TR N: 12) are 50+ years old. This implies that more senior colleagues replied. More than 80% of the professional translators (PR N: 55), 80% seem to be between 30 and 60 years old.



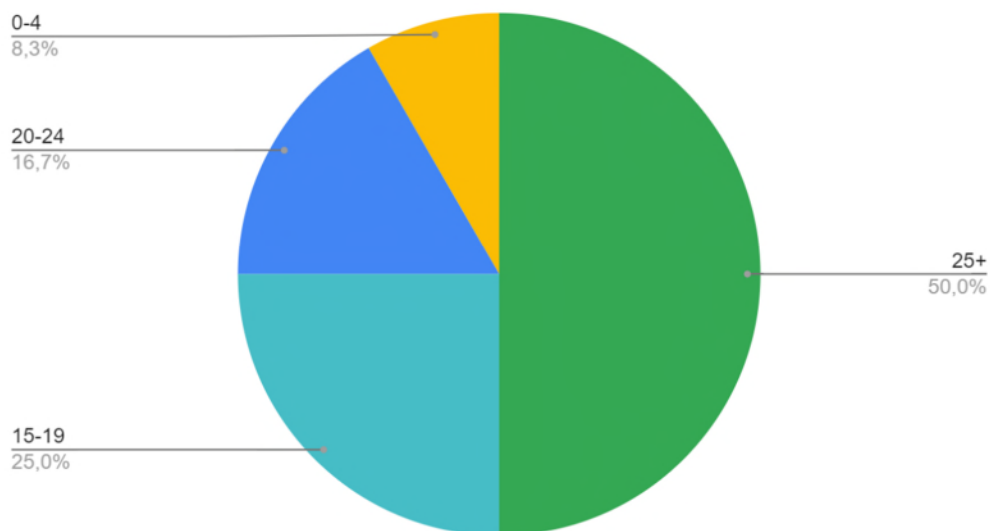
Age - Professionals



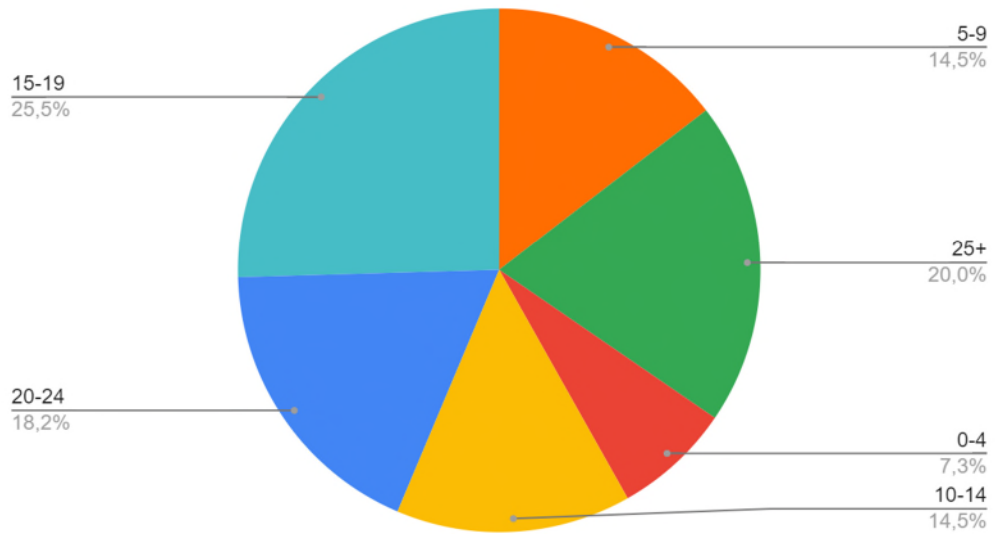
Tables 1 (TR) and 2 (PR). Age of respondents

Regarding professional experience, mostly seasoned trainers replied: most of them with 25+ years of experience. We have more varied groupings as regards translation professionals.

Years of professional experience - Trainers



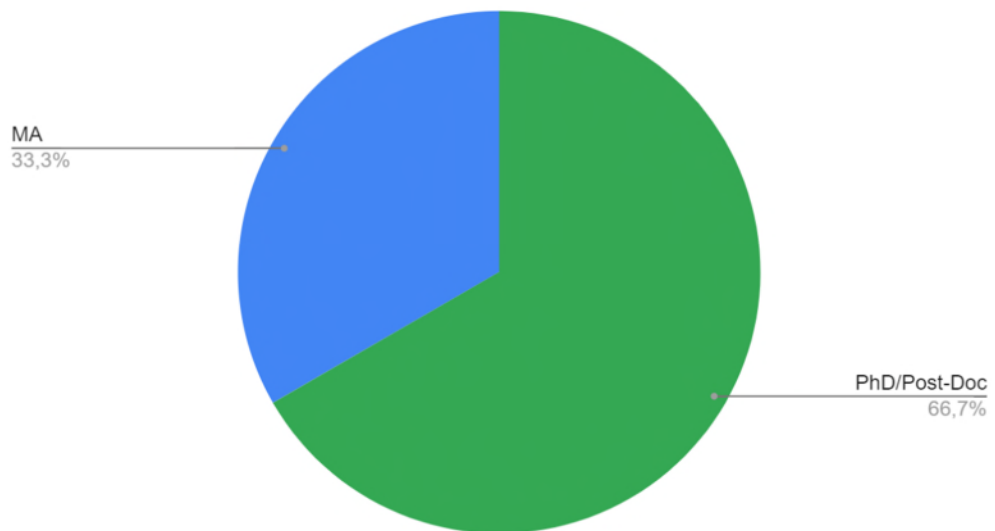
Years of professional experience - Professionals

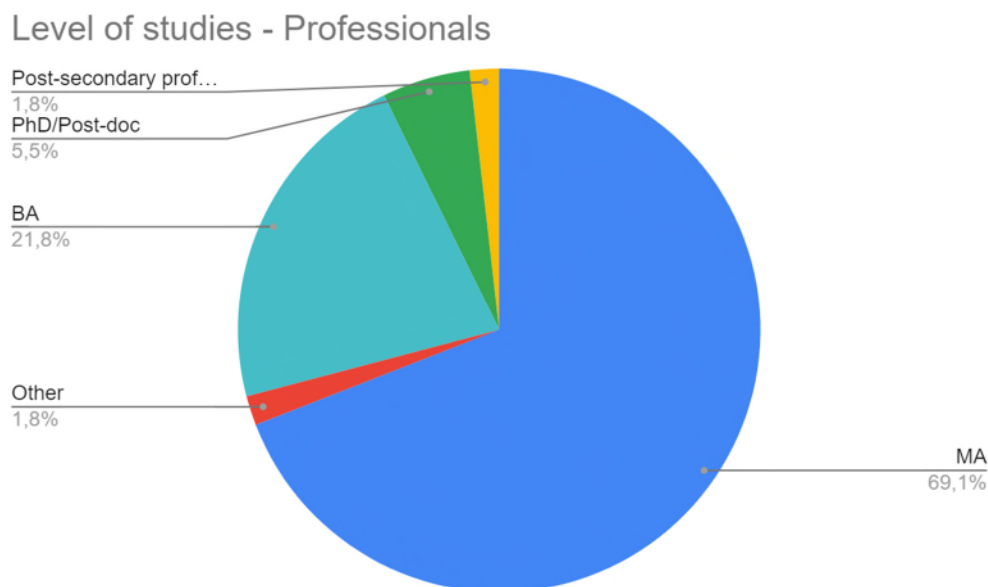


Tables 3 and 4. Years of professional experience

As far as level of studies is concerned, trainers, as expected, are mostly PhD and postdoc holders. Professionals seem to be mostly MA holders.

Level of studies - Trainers





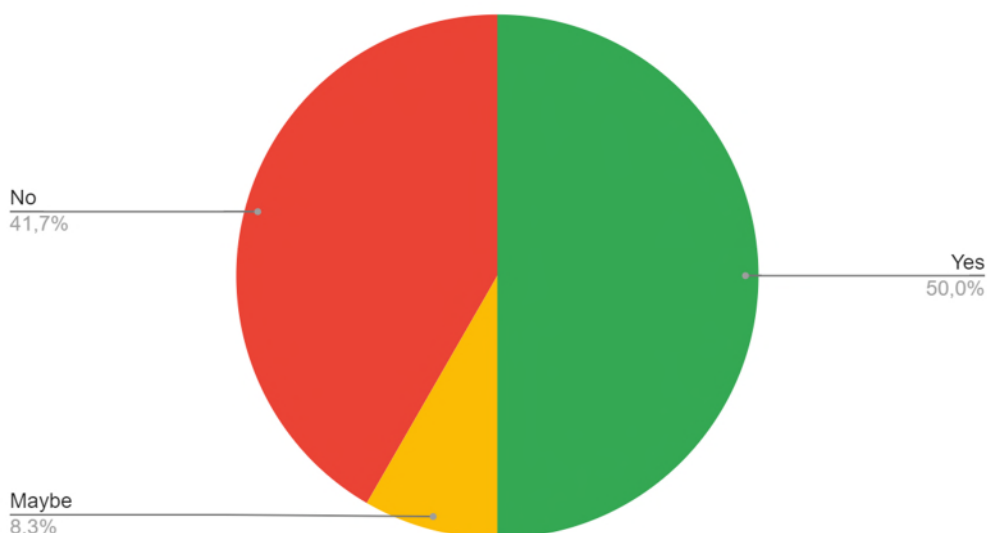
Tables 5 (TR) and 6 (PR). Level of studies

Greek is the mother tongue for most trainers and translation professionals. Both groups teach/translate from English into Greek, while many work from 2 languages into Greek. 8 out of 12 trainers work at universities and 3 out of these 8 as full professors, while the remaining 4 teach in post-secondary institutions such as lifelong learning state or private institutions. Professionals mainly work as freelancers (58.2%), 27.3% have their personal enterprise, while 9.1% work in-house. Finally, as far as their working base is concerned, 11 trainers are based in Greece and 1 in Cyprus; 50 professionals are based in Greece and 5 in Cyprus.

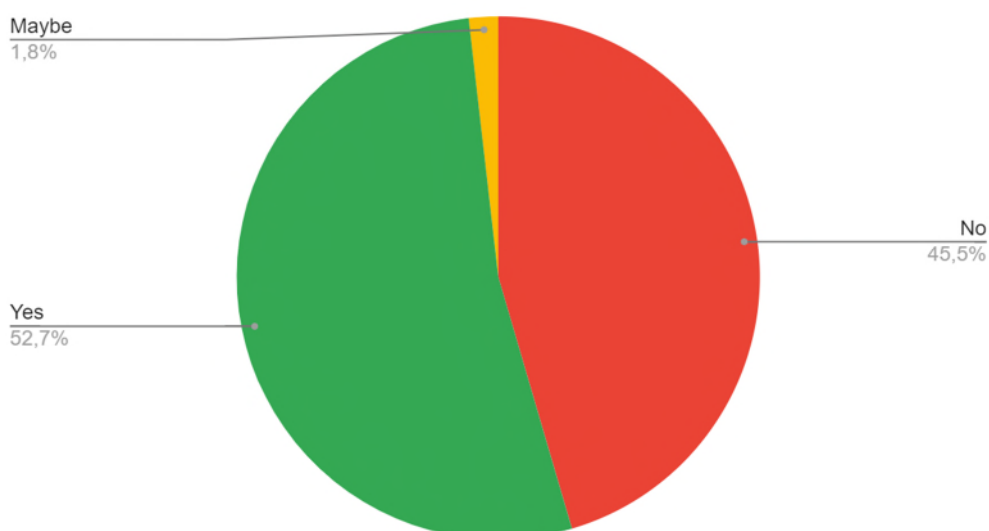
2.2 On ChatGPT

Both groups were asked if they had ever used ChatGPT or other GenAI tools. A “maybe” answer was included into the survey because we noticed, during informal talks, that several colleagues had used Generative AI tools without really being aware of it. Approximately half in both groups have indeed used ChatGPT or other Generative AI tools.

Have you used ChatGPT or other GenAI tools? - Trainers



Have you used ChatGPT or other GenAI tools? - Professionals



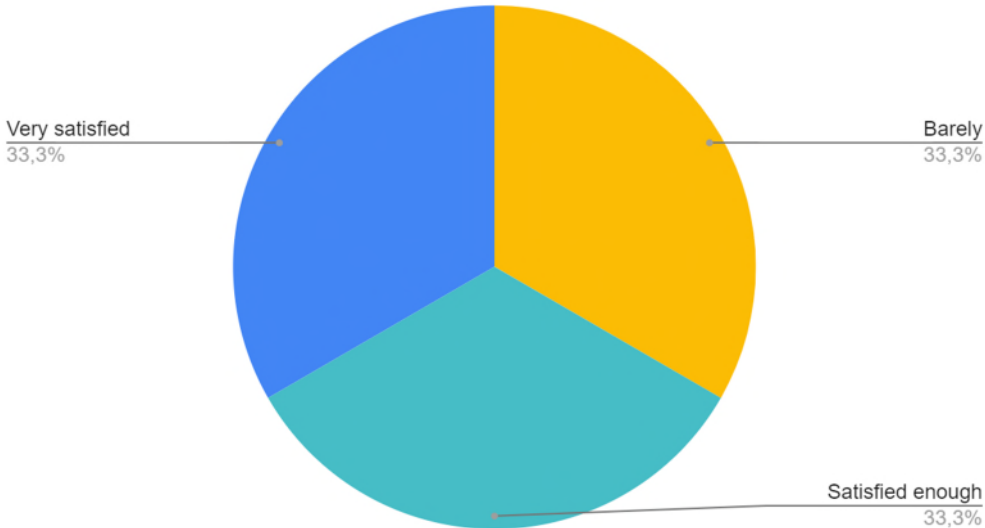
Tables 7 (TR) and 8 (PR). Use of ChatGPT and other GenAI tools.

Trainers who answered positively (N:6), reported that they used ChatGPT to translate a brief text to test the tool; to create definitions; for post-editing in class; to create a summary; to help understand a complex text. Professionals who replied positively (N: 29), reported a greater variety of tasks which also included testing and understanding how ChatGPT works and both translation proper and non-core translation tasks: to improve texts; for creative translation; to summarise; for revision, editing, translation to test the tool; for searches; to see how it works; for copywriting; to draft messages; to create scripts for the automation of language tasks; to

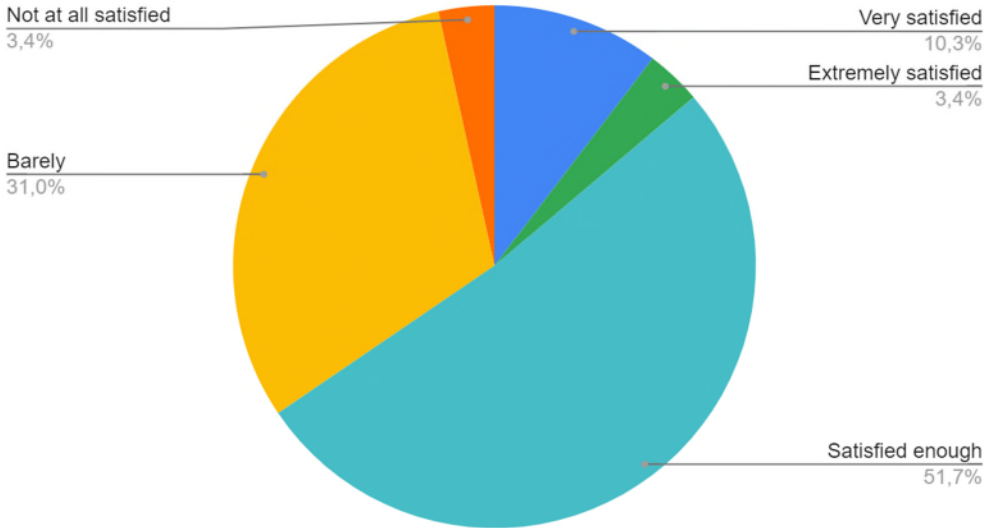
find ideas for argument development; to translate, to search for terminology. One professional also mentioned “for entertainment”.

Trainers are split in 3 groups when reporting their level of satisfaction with ChatGPT: 2 were barely satisfied, 2 satisfied enough, and 2 very satisfied. The responses of professionals are quite different: 10.3% is extremely satisfied, half of them (51.7%) are satisfied enough, while 31% reported barely satisfied.

Level of satisfaction when using ChatGPT - Trainers



Level of satisfaction when using ChatGPT - Professionals



Tables 9 (TR) and 10 (PR). Level of satisfaction when using ChatGPT.

Those respondents who had replied that they had not used ChatGPT (5 trainers, 25 professionals) were asked the reason(s) why. It is interesting that in both groups variations of “I didn’t need to” were reported (1 out of 5 trainers and 10 out 25 professionals). Another interesting finding is reported lack of familiarity with the tool for the time being, (1 out of 5 trainers and 5 out 25 professionals, mainly due to lack of time) and which perhaps might work as an argument for tailored training.

Other replies by trainers who had replied that they had not used ChatGPT included: “It is neither useful nor suitable for my work”; “I trust people more”; “I haven’t checked yet whether it can be effectively applied in teaching”. Replies by professionals who had stated that they had not used ChatGPT included: “It is not integrated in the machine translation tool I am using”; “It does not measure up to the complexity and polysemy of the texts I translate”; “I do not think the level is sufficient”; “Low translation quality, privacy issues”; “I do not think it is honest; the human brain is more valuable”; “I do not think it can help with literary translation”; “I don’t trust it/them”; “I don’t like it”; “I do not see the reason”; “I do not wish so for the time being”.

The one trainer, who replied “I don’t know”, explained that “I have used tools integrating “artificial intelligence” but I lack the necessary technical knowledge to design a course based on making use of artificial intelligence”. The one professional, who replied “I don’t know”, explained that «I am not aware if [it] has been integrated in a machine translation tool I use for my clients”.

Both groups were also asked whether they intend to use ChatGPT or other Generative AI tools in the future and how. Half of the trainers (N: 6/12) answered the question and stated that they would use ChatGPT or other Generative AI tools for a variety of tasks, including teaching and research. More than half of the professionals answered the question (N: 30/55): 4 out of 5 replied positively (24/30), while 1 out of 5 professionals replied negatively and 1 professional has not decided yet. Professionals who are willing to use ChatGPT or other Generative AI tools in the future seem to have the whole process of translation in mind and many want to use such tools as a springboard for ideas as is evident from the grouping in the table below.

Trainers	Professionals
<ul style="list-style-type: none"> • For bibliographic information • To help me understand some obscure pieces in English • Yes, for research purposes and text editing • Only for draft translation • By incorporating them in translation courses • Yes, in class and during exams 	<ul style="list-style-type: none"> • For entertainment purposes • For initial inspiration and ideas • For developing arguments and managing difficult situations • For knowledge questions • For documentation • For term mining and searches • For clarifying meaning, • For summarising • To create first [translation] draft

	<ul style="list-style-type: none"> • For improving texts (proofreading, editing), copywriting • For a quick quality check or for alignment • For saving time but not [...] under good deadlines • Yes, to the extent that quality of the tool(s) is judged to be satisfactory for the execution of daily tasks. • Yes, within reason • Yes, depending on the capabilities in the future
--	---

Table 311. Intention to use ChatGPT and other GenAI tools in the future (selection).

It is interesting that 2 trainers who had replied earlier that they were barely or not at all satisfied with ChatGPT nevertheless answered that they would use ChatGPT for draft translations or even incorporate it in their classes. Replies by professionals who had replied earlier that they were barely or not at all satisfied with ChatGPT are quite diverse: responses range from absolute “no”/“not for the time being, because I do not need it”/“no, unless it is greatly improved” to “I don’t know” to “for saving time but not [...] when deadlines are good” to “for knowledge questions” and even “for the first translation] draft”.

2.3 The future

Both groups were asked to comment in one phrase how they see the future of translator training or the translation sector, respectively, in relation to ChatGPT and other GenAI tools. Replies vary from positive comments to negative comments in both groups, with greater “granularity” in professionals, also due to the higher number of respondents. Similar answers are grouped together and presented from the positive to the negative end of the spectrum for ease of reference in the table below.

Trainers	Professionals
<ul style="list-style-type: none"> • Positively, provided their use is controlled. The profession is not under threat for the time being • Tools should be integrated in translator training programmes 	<ul style="list-style-type: none"> • Very positively • Normal/Fine/Easier/Optimistic • Harmonious collaboration • Useful aid for translators • If it makes our job easier, it’s good

<ul style="list-style-type: none"> • Learning how to ethically use them is sine qua non • Translator training will be revolutionized by knowledge from Cognitive Psychology • There is a tendency toward promoting post-editing • It is too early to say • They may be used as a tool but the intervention by professionals will be necessary • The need to stay constantly up-to-date will be strong • Unsettling 	<ul style="list-style-type: none"> • Use of AI by professionals • Not necessarily pessimistic • I don't know • Uncertain (future) • In flux • AI makes our daily lives easier but it should be used with caution • The translator will become a post editing language specialist • It will be imposed by companies • We are already observing a practice where clients offer lower prices citing ChatGPT as an excuse • If machines can do everything people can, then we have a problem • Scary/Dark/Ominous
--	--

Table 322. The future in relation to ChatGPT and other GenAI tools (selection).

3 Concluding remarks

The number of respondents was lower than one might expect. Measures were taken so that the questionnaires reached the most relevant people. Methodologically speaking, however, blind questionnaires are susceptible to low response rate (Ornstein 2013), and that diminishes the reliability of the data, since the responses collected represent the views of the sample at hand and may not extrapolate to the population under study. The general nature of questionnaire items also contributed to an overview rather than a detailed view.

The current sample of self-reports thus offers a glimpse into what is happening in Greece and Cyprus. Half of the respondents have used ChatGPT for one purpose or another and it would be interesting to check whether respondents of the 2023 questionnaire still have the same ideas about and attitude toward ChatGPT and, pedagogically speaking, it would be worthwhile looking into which tasks led to higher levels of satisfaction. From another perspective, it would also be interesting to see what would happen if those who are not yet familiar with it either due to lack of time or due to their particular circumstances did get a chance to use the tool and whether such cases might benefit from specific and structured training approaches; both trainers who might feel more confident toward designing more relatable courses or introducing relevant components and professionals who might use the tool to optimise their translation work. Perhaps a more systematic mapping of the situation in many countries is called for. Such mapping would contribute toward structured familiarisation and relevant interventions based on collaboration between academia, the professionals as well as large industry stakeholders

might lead to a clearer and more realistic view of ChatGPT (and, possibly, other GenerativeAI tools), its inherent “mutating” nature, its advantages, drawbacks and dangers, and, consequently, toward more conscious decision-making about its use, a more confident, future-facing attitude and, hopefully, a beneficial symbiosis.

Acknowledgements

I would like to thank all respondents as well as the Panhellenic Association of Translators, the Panhellenic Association of Professional Translators Graduates of the Ionian University, the Association of Translators-Editors-Proofreaders and the Pancyprian Union of Graduate Translators & Interpreters for their support.

References

Ornstein, Michael. 2013. *A companion to survey research*. SAGE Publications Ltd, <https://doi.org/10.4135/9781473913943>